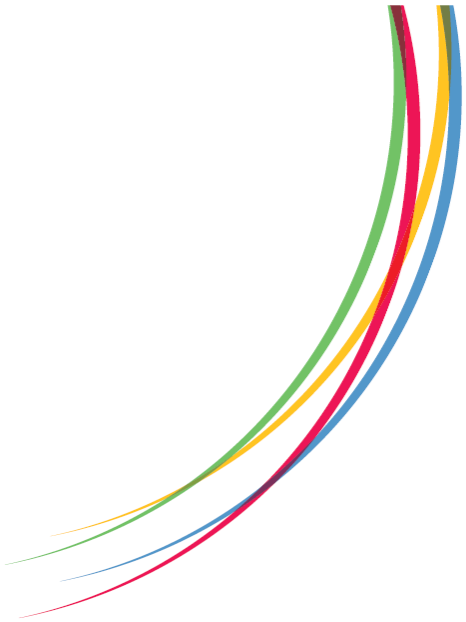


Understanding Society
The UK Household Longitudinal Study
Waves 1-4,
User Manual

Edited by
Gundi Knies
Institute for Social and Economic Research
University of Essex
Colchester
Essex

Version 1.1, October 2014

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

CONTENTS

Table of Tables 3
Table of Figures 4
List of Abbreviations 5
1. Introduction 6
1.1. What is *Understanding Society*? 6
1.2. How to Navigate this User Manual 6
2. *Understanding Society* Study Design 7
2.1. Overview 7
2.2. Sample Design 8
2.2.1. General Population Sample 8
2.2.2. General Population Comparison Sample 9
2.2.3. Ethnic Minority Boost Sample 9
2.2.4. Former BHPS Sample 10
2.2.5. Sample Status and Following Rules 10
2.3. Data Collection and Response Outcomes 11
2.3.1. Overview 11
2.3.2. Data Collection 12
2.3.3. Panel Membership and Panel Maintenance 15
2.3.4. Response Outcomes 16
2.4. Data Processing and Cleaning 29
2.4.1. Coding 29
2.5. Documentation of the Survey Instruments 30
2.5.1. Reading the Questionnaires 30
2.5.2. Summary of Questionnaire Modules 32
2.5.3. Content Highlights by Wave 37
2.5.4. Changes to the Questionnaire 39
2.5.5. Other Fieldwork Materials 39
3. *Understanding Society* Data 40
3.1. Information About Data Files 40
3.1.1. Paradata 41
3.2. Information About Variables 42
3.2.1. Learning About the Study Variables 42
3.2.2. Variable Naming and Labelling Conventions 42
3.2.3. Variable Values and Labels 43

3.2.4. Identifiers and Other Useful Variables 43
3.3. Notes on Using the BHPS Sample 45
3.4. Notes on Using the “Extra 5 minutes” questions 46
3.5. Derived Variables 46
3.6. Sample Design Variables and Analysis 48
3.6.1. Primary Sampling Units (PSUs) 48
3.6.2. Strata 49
3.6.3. Analysis Example Using Stata 49
3.7. Weighting Adjustments for the Wave 4 Release 49
3.7.1. Selecting the Correct Weight for Your Analysis 50
3.7.2. Naming Conventions for Weighting Variables 57
3.7.3. Technical Details 58
3.8. Imputation of Income Variables 72
3.8.1. What Do We Impute? 73
3.8.2. Imputation Procedures 74
3.8.3. Item nonresponse on income variables in the individual questionnaire 77
3.8.4. Item Nonresponse for Income Variables in the Proxy Questionnaire .. 78
3.8.5. Individual Non-respondents Without Proxy Questionnaire 78
3.8.6. Net Income Estimates 79
3.9. Example Code for Matching Files 80
3.10. Preserving Confidentiality 83
4. Data Access 84
4.1. Release Versions 84
4.1.1. End User Licence (EUL) Data 84
4.1.2. Special Licence (SL) Data 84
4.1.3. Secure Access 85
4.2. Revisions to Previous Releases 85
4.3. Links to other studies in the UKHLS family 86
4.3.1. The British Household Panel Survey 86
4.3.2. The Waves 2-3 Nurse Health Assessment 86
4.3.3. The UKHLS Innovation Panel 86
5. Citations and Acknowledgements 86
5.1. Citation of the Data 86
5.2. Citation of the User Manual 86
5.3. Acknowledgments 87
6. References 88

TABLE OF TABLES

Table 1: Timing of data collection start 12
Table 2: Household response rates among eligible households, Wave 1 16
Table 3: Individual response rates, Wave 1 17
Table 4: Household response rates, Wave 2* 18
Table 5: Wave 2 Cross-sectional individual adult response rates by sample origin 20
Table 6: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at Wave 1 21
Table 7: Household response rates, Wave 3* 22
Table 8: Wave 3 cross-sectional individual adult response rates by sample origin 24
Table 9: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at Wave 2 25
Table 10: Household response rates, Wave 4* 26
Table 11: Wave 4 Cross-sectional individual adult response rates by sample origin 27
Table 12: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at Wave 3 28
Table 13: Summary of questionnaire modules 32
Table 14: Summary of adult self-completion content 36
Table 15: List of select data files: Data from responding sample members 40
Table 16: List of select data files: Data from enumerated sample members 41
Table 17: List of select data files: Cross-wave files 41
Table 18: List of select data files: Paradata 42
Table 19: Missing value codes 43
Table 20: Some useful variables 44
Table 21: Description of the UKHLS Primary Sampling Unit variable (*w*_psu) 48
Table 22: Description of the UKHLS stratification variable (*w*_strata) 49
Table 23: Selecting the correct weight: Hierarchy of analysis levels 50
Table 24: Weight variables for analyses using household grid or household interview 52
Table 25: Weights for analysis using adult main and proxy interviews 53
Table 26: Weights for analysis using adult main interviews 54
Table 27: Weights for analysis using adult “Extra 5 minutes” interview 54
Table 28: Weights for analysis using adult self-completion 54
Table 29: Weights for analysis using youth self-completion 55
Table 30: Weights for analysis using nurse health assessment data 56
Table 31: Design and inclusion weights* 56

TABLE OF FIGURES

Figure 1: Mark-up of household questionnaire 31
Figure 2: Mark-up of question with looping from individual questionnaire 31
Figure 3: Stata Code: Distributing household level information to individuals 81
Figure 4: Stata Code: Summarizing individual level information to household level. 81
Figure 5: Stata Code: Matching individuals within a household 82
Figure 6: Stata Code: Using the egoalt file to create household composition variables 82
Figure 7: Stata Code: Merging individual files across waves into long format..... 83
Figure 8: Stata Code: Merging individual files across waves into wide format 83

LIST OF ABBREVIATIONS

BHPS	British Household Panel Survey
HBAI	Households Below Average Income Statistics
CAPI	Computer Assisted Personal Interview
CASCOT	Computer Assisted Structured Coding Tool
CASI	Computer Assisted Self-Interview
DWP	Department for Work and Pensions
ECHP	European Community Household Panel
EMB	Ethnic Minority Boost
ESRC	Economic and Social Research Council
GOR	General Office Region
GPC	General Population Comparison
GPS	General Population Sample
ICE	Imputation by chained equations
IP	UKHLS Innovation Panel
ISER	Institute for Social and Economic Research
LDA	Low density ethnic minority area
LS	Little and Su method
LSOA	Lower Layer Super Output Area
MMU	Multi-Mode Unit
MSOA	Middle Layer Super Output Area
NatCen	National Centre for Social Research
NE	New entrant to the study
NIHPS	Northern Ireland Household Panel Survey
NISRA	Northern Ireland Statistics and Research Agency
ONS	Office for National Statistics
OSM	Original Sample Member
PC	Population Carryover method
PMM	Predictive mean matching
PSM	Permanent Sample Member
SIC	ONS Standard Industry Code
SOA	Super Output Area
SOC	ONS Standard Occupational Classification
TSM	Temporary Sample Member
UKHLS	UK Household Longitudinal Study (official acronym for *Understanding Society*)

1. INTRODUCTION

1.1. WHAT IS *UNDERSTANDING SOCIETY*?

Understanding Society, the UK Household Longitudinal Study (UKHLS), is a longitudinal survey of the members of approximately 40,000 households in the United Kingdom, i.e., the geographical area of the countries England, Scotland, Wales and Northern Ireland. Households recruited at the first round of data collection are visited each year to collect information on changes to their household and individual circumstances. Interviews are typically carried out face-to-face in respondents' homes by trained interviewers. Data collection for each wave takes place over a 24 month period as shown in Figure 1. Note that the periods of waves overlap, and that individual respondents are interviewed around the same time each year.

Understanding Society is funded by the Economic and Social Research Council (ESRC) and with funding from multiple government departments (the Department for Work and Pensions (DWP), the Department for Education, the Department for Transport, the Department for Culture, Media and Sport, the Department for Communities and Local Government, the Department of Health, the Scottish Government, the Welsh Assembly Government, the Northern Ireland Executive, the Department for Environment, Food and Rural Affairs, and the Food Standards Agency). The scientific leadership team is from the Institute for Social and Economic Research (ISER) of the University of Essex, the University of Warwick, and the London School of Economics. Professor Nick Buck is the principal investigator. Fieldwork for waves 1 to 4 was conducted by the National Centre for Social Research (NatCen) with collaboration with the Central Survey Unit of the Northern Ireland Statistics and Research Agency (NISRA) in Northern Ireland.

The overall purpose of *Understanding Society* is to provide high quality longitudinal data about subjects such as health, work, education, income, family, and social life to help understand the long term effects of social and economic change, as well as policy interventions designed to impact upon the general well-being of the UK population. To this end, the study collects both objective and subjective indicators and offers opportunities for research within and across multiple disciplines such as sociology and economics, geography, psychology and health sciences. The study also provides a platform for additional data collections.

1.2. HOW TO NAVIGATE THIS USER MANUAL

This release has data for the *Understanding Society* Main Survey which collects information from the UK General Population Sample (GPS) and the Ethnic Minority Boost (EMB) sample. From Wave 2 onward the Main Survey also includes information collected from continuing participants of the British Household Panel Survey (BHPS), a household panel survey of around 8,000 households in the UK, which has completed 18 annual waves of data collection and has been run by ISER since it began in 1991. To learn more about the BHPS and other components of *Understanding Society*, see Section 4.3, below.

The User Manual is structured as follows. We first present the general aspects of the study design (Section 2) which covers sample design (Section 2.2), data collection (Section 2.3), data processing (Section 2.4), and questionnaire content (Section 2.5). Section 3 provides a description of UKHLS data files and variables, covering derived variables (Section 3.5), weighting adjustments (Section 3.7), imputation of income (Section 3.8) and example code for matching information contained in different files (Section 3.8). Information on how to access the data is provided in Section 4. Section 5 provides additional information about further studies in the *Understanding Society* family: currently the BHPS, the UKHLS Nurse Health Assessment which took place in Waves 2 and Wave 3, and the UKHLS Innovation Panel.

As an introduction to the UKHLS Main Survey data and documentation we particularly recommend the following reading:

- The summary of the general questionnaire content (Section 2.5), and notes on naming conventions (Section 3.2.2),
- The sections on sample design (Section 2.2), weighting adjustments (Section 3.7) and data collection and response outcomes (Section 2.3).
- Variable level descriptions of the data can be found on the study website (https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation). The online documentation has extensive links between questions and detailed views of variables and data files. There is also a search facility for searching questions, variables, modules, and data files.
- The example Stata code for matching variables from different data files (Section 3.8).

In assembling the documentation, we have drawn upon the documentation for the BHPS, see Taylor (2010) and http://www.iser.essex.ac.uk/bhps.

2. *UNDERSTANDING SOCIETY* STUDY DESIGN

2.1. OVERVIEW

Understanding Society is a panel survey of households with yearly interviews. Data collection for a single wave is scheduled across 24 months. The study began with a representative probability sample of households. There is an extended discussion of sample design in Section 2.2, and in Lynn (2009). Adult household members (age 16 or older) are interviewed and the same individuals are re-interviewed in successive years to see how things have changed. Household members aged 10-15 years are asked to complete a short self-completion youth questionnaire. Children become eligible for a full interview once they reach the age of 16. We refer to them as “rising 16’s”.

The overall study has multiple sample components. In the Main Survey there is the a) General Population Sample (GPS), with its subset the General Population Comparison (GPC) sample, b) the Ethnic Minority Boost (EMB) sample, and c) participants from the BHPS. The instruments for the first three components are the same except the EMB sample and the GPC sample have an “Extra 5 minutes” of questions specifically relevant to ethnic minority communities (e.g., ethnic identity and remittances).

In Waves 2 and 3, *Understanding Society* augmented survey questions with direct health assessments and the collection of blood samples. The Health Assessment data can be accessed through the UK Data Service, SN7251.

In addition, there is a separate survey, the Innovation Panel (IP), which is fielded in the year before the main survey. It tests varying measurement issues, and its instruments are somewhat different from the main survey. The IP can be accessed through the UK Data Service, SN 6849.

2.2. SAMPLE DESIGN

The *Understanding Society* sample consists of a new large General Population Sample (GPS) plus four other components (the Ethnic Minority Boost (EMB) sample, the General Population Comparison (GPC) sample, the former BHPS sample and the UK Innovation Panel (IP) sample). The design of all five components is described in more detail in an *Understanding Society* working paper, see Lynn (2009). The Innovation Panel is prepared as a separate study, which can be accessed via the UK Data Service, SN 6849. The GPS is based upon two separate samples of residential addresses of England, Scotland and Wales and for Northern Ireland. The England, Scotland and Wales sample is a proportionately stratified (equal probability), clustered sample of addresses selected from the Postcode Address File. Northern Ireland has an unclustered systematic random sample of addresses selected from the Land and Property Services Agency list of domestic addresses.

2.2.1. GENERAL POPULATION SAMPLE

The sample for England, Scotland and Wales was selected in two stages. The first stage was to select a sample of postcode sectors as the primary sampling units (PSU's). The second stage was to select addresses within each sampled sector. Prior to selection, any postcode sector with fewer than 500 residential addresses was grouped with an adjacent sector and thereafter treated as a single sector. The list of all sectors was then sorted into twelve geographical strata, consisting of ten regions in England plus Scotland and Wales as separate strata. Within each of the twelve strata, sectors were sorted into three sub-strata based upon the proportion of household reference persons classified as non-manual workers, from 2001 Census data. Within each of the 36 sub-strata, sectors were then sorted into three further sub-divisions based on population density (households per hectare) and within each of the 108 resultant sub-divisions, sectors were listed in order of ethnic minority density. From the sorted list, a systematic random sample of 2,640 sectors was selected, with probability proportional to the number of residential addresses in the sector. These sectors were then allocated systematically to 24 monthly samples, with 110 sectors in each monthly sample. Within each postal sector, 18 addresses were selected using systematic random sampling. The England, Scotland and Wales sample in this data release is based upon an initial sample of 47,520 addresses.

In Northern Ireland, 2,395 addresses were selected in a single stage from the list of domestic addresses. In combination, this data release is therefore based upon a total of 49,915 addresses.

At each address, the final stage of sampling was carried out by field interviewers. This consisted of identifying persons to be defined as sample members. All persons resident at each sample address at the time the interviewer made contact were deemed to be a sample member, with the exception of the small proportion of

addresses that contained more than three dwellings or households. In those cases, three dwellings or households were sub-sampled at random.

2.2.2. GENERAL POPULATION COMPARISON SAMPLE

The General Population Comparison sample (GPC) has one sampled address for 40% of the selected postal sectors in General Population Sample (GPS) component for Great Britain. In other words, of the 2,640 general population sectors, 60% of them (1,584) contain 18 GPS addresses and the other 40% contain 17 GPS addresses and one GPC address. The persons in these households will be designated as members of the GPC sample, regardless of ethnic group membership. Members of the GPC sample are a random subsample of the GPS component and they should be included in analyses of the GPS component.

2.2.3. ETHNIC MINORITY BOOST SAMPLE

The EMB sample was designed to provide at least 1,000 adults from each of five groups: Indian, Pakistani, Bangladeshi, Caribbean, and African.

The initial step was identifying postal sectors with relatively high proportions of relevant ethnic minority groups, based upon 2001 Census data and more recent Annual Population Survey data. The set of 3,145 sectors constituted approximately 35% of the sectors in Great Britain and covered between 82% and 93% of the population of the five ethnic minority groups.

The 3,145 sectors were sorted into four strata based on the expected number of ethnic minority households that would be identified by the sampling and screening procedures (see Berthoud et al., 2009 for details). All sectors were included for the stratum where a yield of three or more households was expected. In the other three strata, sectors were sub-sampled at rates of 1 in 4, 1 in 8, or 1 in 16 respectively. This was done to constrain the number of sectors that might have just one or two eligible sample households (or even none). The total number of postal sectors selected for inclusion in the EMB sample was 771. Of these 6 were in Scotland, 7 were in Wales, and the remaining 758 were in England, with a concentration in London (412 sectors).

The number of addresses selected per postal sector ranged from 15 to 103. Sampling fractions varied across the sectors in a way designed to deliver target numbers of respondents in each target ethnic minority group with adequate statistical efficiency (see Berthoud et al., 2009 for more details). In sectors selected for both the GPS component and the EMB sample, a single systematic sample of the required total number of addresses was selected and allocated in a systematic way to the two sample components, thus ensuring that both sample components are spread throughout the whole sector.

The final stage of sampling was done by the interviewers. The steps are described in the Project Instructions for Interviewers. At addresses containing more than three dwellings or households, the procedures to sub-select dwellings or households were as described above for the GPS component. Within each household, rather than all resident persons becoming sample members, there were three additional steps:

- A “screen” was carried out to identify whether there were any persons from target ethnic groups in the household.

- A random mechanism was applied to certain target groups identified by the screen in order to select only a desired proportion into the sample (non-mixed Indian, Pakistani, non-mixed Caribbean, African, Far Eastern, Middle Eastern). For other target groups, all resident persons were included in the sample (mixed Indian, Bangladeshi, mixed Caribbean, Sri Lankan, Chinese, Turkish).
- In households included in the sample in the previous two steps, all members of target ethnic groups were deemed to be members of the EMB sample (including children). All persons of other ethnic groups are not EMB sample members. They will be interviewed as temporary sample members for so long as they remain co-resident with at least one EMB sample member.

The overall sampling fractions combine a) the probability of sampling the sector, b) the fraction of addresses selected within the sector, and c) the probability of a household being retained following the application of the random selection mechanism described above.

2.2.4. FORMER BHPS SAMPLE

The sample issued at Wave 2 consisted of all members from the BHPS sample who were still active at Wave 18 of the BHPS and who had not refused consent to be issued as part of the *Understanding Society* sample. It should be noted that the BHPS sample contains different components, including the original sample (first selected in1991), boost samples in Scotland and Wales (first selected in 1999), and a Northern Ireland sample (selected in 2001). For further details of the BHPS sample, see section IV of the BHPS User Guide (http://www.iser.essex.ac.uk/bhps/documentation/vola/vola.html).

2.2.5. SAMPLE STATUS AND FOLLOWING RULES

There are three possible sample statuses: Original Sample Members (OSMs), Temporary Sample Members (TSMs), and Permanent Sample members (PSMs). The definitions are as follows:

2.2.5.1. Original Sample Members (OSMs)

All members of *Understanding Society* General Population Sample households enumerated at Wave 1, including absent household members and those living in institutions who would otherwise be resident, are Original Sample Members (OSMs). All ethnic minority members of an enumerated household eligible for inclusion in the EMB sample are OSMs. In all of these samples, any child born to an OSM mother after Wave 1 and observed to be co-resident with the mother at the survey wave following the child's birth is an OSM. In the former BHPS sample, OSMs are those who were enumerated at the first wave of the sample from which they come (Wave 1 for the original sample, Wave 9 for the Scotland and Wales boost samples, Wave 11 for Northern Ireland) or who were subsequently born to an OSM mother or father (or both). From Wave 2 onwards of *Understanding Society*, in the former BHPS sample, as for the rest of the *Understanding Society* sample, only children born to an OSM mother will themselves become an OSM. OSMs, of all ages, are followed for interview and remain eligible as long as they are resident within the UK. They remain potentially eligible sample members for the life of survey.

The case may arise where the only OSM in the household is a child. Other household members are then TSMs so long as they are co-resident with the child, and therefore eligible for interview, even if the child is not yet old enough to be eligible for interview. If the OSM child moves house, they are followed to their new address and those living with the OSM child are eligible for interview. If the OSM child moves into an institution, where normally just the OSM/PSM would be interviewed and not co-residents, a split-off household is created containing only the OSM child and the household enumeration grid completed. The child OSM is an eligible sample member, even if they are not eligible for interview because of their age.

2.2.5.2. Temporary Sample Members (TSMs)

Any members of an enumerated household eligible for inclusion in the EMB sample at Wave 1 who are not from a qualifying ethnic minority are Temporary Sample Members (TSMs) at Wave 1. This was the only category of TSM at Wave 1. In all parts of the sample, any new person found to be co-resident in an OSM or PSM household after Wave 1 is a TSM. This would include any child born to an OSM father after Wave 1 but not an OSM mother and observed to be co-resident with the father (or any other OSM) at the survey wave following the child's birth. TSMs remain eligible for interview as long as co-resident in an OSM/PSM household. TSMs who are not co-resident in an OSM/PSM household are not followed and become ineligible for interview. TSMs are identified as re-joiners if they are subsequently found in an OSM/PSM household and then become eligible for interview.

2.2.5.3. Permanent Sample Members (PSMs)

PSMs are TSMs who are followed for interview after they no longer live with an OSM. This is done for substantive research reasons because of the additional contextual information they may provide for the analysis of OSMs. At present, there is only one category of PSM, but others may be defined in the future. Any TSM father of an OSM child born after Wave 1 and observed to be co-resident with the child at the survey wave following the child's birth is a PSM. PSMs remain potentially eligible for interview for the life of survey.

2.3. Data Collection and Response Outcomes

2.3.1. Overview

The UKHLS is issued to field as 24 monthly samples. There is some variation in this pattern. The Northern Ireland and the former BHPS sample components are issued over the first 12 months of the wave. Table 1 shows the timing of the sample issue for the data included in this release.

Most of the data collection is conducted face-to-face via computer aided personal interview (CAPI). There are also self-completion instruments for youth and adults. The youth instruments are administered on paper. The adult self-completion questionnaire shifted from paper to computer administered self- interview (CASI) in Wave 3. From Wave 3 onwards, there was also a telephone mop-up at the end of the fieldwork period for each sample month.

Table 1: Timing of data collection start

Year	Quarter	Survey
2008	Q1	
	Q2	
	Q3	BHPS Wave 18
	Q4	
2009	Q1	Wave 1, year 1
	Q2	
	Q3	
	Q4	
2010	Q1	Wave 1, year 2
	Q2	
	Q3	
	Q4	
2011	Q1	Wave 2, year 1*
	Q2	
	Q3	
	Q4	
2012	Q1	Wave 2, year 2
	Q2	
	Q3	
	Q4	
2013	Q1	Wave 3, year 1
	Q2	
	Q3	
	Q4	
		Wave 3, year 2
		Wave 4, year 1
		Wave 4, year 2

* BHPS becomes sample component in Wave 2, year 1

2.3.2. DATA COLLECTION

2.3.2.1. The players, who does what

ISER, together with NatCen and the Central Survey Unit of NISRA, work closely together on all aspects of data collection, implementing an agreed set of survey procedures designed to ensure adequate response and effective data quality.

ISER has the primary responsibility for design work. NatCen manages fieldwork, editing, and coding and data-entry. It also advises on the design of all research instruments. NISRA collaborates with NatCen and is responsible for fieldwork in Northern Ireland. ISER plays a major role in quality control through specification of fieldwork practices; survey materials; editing and coding requirements; and monitoring and analysing weekly fieldwork progress reports. This working relationship is reinforced by an agreed set of survey-specific procedures to ensure adequate response and effective data quality. Full details of these, and other technical aspects of the data collection and fieldwork, coding, and data processing are found in the *Technical Reports,* published each wave on the *Understanding*

Society website (see https://www.understandingsociety.ac.uk/documentation/mainstage/technical-reports).

2.3.2.2. Getting Ready for Fieldwork

Prior to the first wave of the UKHLS Main Survey, there were two small pilot studies and a dress rehearsal. A cognitive pilot of 70 individuals was conducted March – April 2008 to test screening and other questions relevant to the ethnicity strand. A translation pilot was conducted in June 2008: 50 interviews were carried out using Bengali and Punjabi translations of the questionnaire to see if there were problems with the operation of the translation program or problems with interviewing with the translated instruments. A run-through of all data collection instruments and procedures in 100 households, called a dress rehearsal, took place August-September 2008.

A pilot or run-in for Wave 2 tested all instruments and data collection procedures. For this wave, the data collection also focused on assessing any problems with integrating members of the former BHPS sample component, which includes a small segment conducted by telephone interviews. In all 237 households were issued. Of these, 91 were households interviewed in the Wave 1 pilot. The BHPS sample component was represented by households that were part of the BHPS between 1997 and 2001, the European Community Household Panel (ECHP). Households for which we had a telephone number were issued to telephone interview to test the telephone interview instruments and procedures. The Wave 2 pilot took place September-October 2009.

The Wave 3 dress rehearsal pilot took place September-November 2010. Interviewers returned to the same sample that was interviewed at the Wave 1 and 2 pilots.

The Wave 4 dress rehearsal pilot took place September-November 2011. Interviewers returned to the same sample that was interviewed at the previous waves' pilots.

2.3.2.3. Interviewers

We have tried to use interviewers of above average levels of experience and ability because of the demanding nature of *Understanding Society***.** The majority of interviewers in Northern Ireland had worked on the BHPS Northern Ireland component (the Northern Ireland Household Panel Survey), and were familiar with the design and operation of *Understanding Society*.

In addition to general interviewer training, interviewers working on the study attended a one day survey-specific briefing. Generally around 12-20 interviewers attended each briefing, along with two or three briefing managers or area managers. The briefings were led by at least one researcher from NatCen with the majority also attended by ISER staff. The briefings in Wave 1 took place across the UK; Belfast, Birmingham, Brentwood, Bristol, Derby, Edinburgh, Glasgow, Leeds, London and Manchester. Similar topics and locations were used for the Wave 2 briefings. At Wave 3, the Edinburgh briefing was dropped and two briefings were held in Glasgow. Additional briefings were added in Bury St. Edmonds, Liverpool and Gateshead.

The morning sessions were devoted to fieldwork procedures, for example the administrative forms to record contact information, and how to deal with the complexities of multiple dwelling units and multiple households. The afternoon was spent discussing the survey content and reviewing and working with the Blaise computer aided personal interview (CAPI) instrument. At Wave 3 there were two types of briefing; for interviewers experienced with the study or for interviewers with experience who were new to the study. The latter briefing went into more detail about the background of *Understanding Society*, early findings, the more technical details of the sample, and the task of enumerating the household.

At Wave 4, the style of briefing changed in Great Britain. Interviewers had worked on the study for three waves and were familiar with the mechanics of how to conduct the survey. For those interviewers returning for Wave 4, the focus of the briefing switched from the survey procedures to motivating the interviewers and giving them information to enable them to motivate the sample members when making contact. Interviewers who were new to the survey still attended a standard briefing, as did interviewers in Northern Ireland. These standard briefings were held in Belfast (3) and London (1).

Experienced interviewers attended ‘conference-style’ briefings. These briefings were much larger than standard briefings, with 150-250 interviewers attending each event. There were three such events held prior to the start of Wave 4; in Birmingham, Liverpool and London. During the breaks in day, there were stalls and displays of media coverage, research findings, and information about the study, a Twitter stand and an area where interviewers could write questions on post-it notes for discussion later in the day. The content of the briefing consisted of ‘plenary’ sessions where an overview of progress on the study “so far” was presented, along with researchers from ISER or the LSE talking about how they used *Understanding Society* in their research, videos of the Chief Executive of NatCen (Penny Young) and the ISER Director (Professor Heather Laurie MBE) were shown, and medals awarded to interviewers who had achieved 100% response rate in any of their allocations at Wave 3. Once during the morning, and once in the afternoon, there were a number of ‘break-out’ sessions with small groups of interviewers to share best practice and experience of (i) contact and co-operation and (ii) how to deal with household splits and allocating outcome codes. The discussions of these break-out sessions were then discussed at the plenary sessions.

Interviewers were assigned to specific areas. For Wave 1, 911 interviewers were employed to cover 3,517 areas in the sample. The number of interviewers briefed in Wave 2 was 819 and 746 at Wave 3. At Wave 4, 692 interviewers worked on the study.

2.3.2.4. Fieldwork

When beginning fieldwork for Wave 1, we did not know who was in the sample. Interviewers mailed an introductory card from ISER to all sampled addresses (addressed to "The Occupier"), together with a small leaflet outlining the purpose of the survey. Then the interviewer called within a week of the mailing. At the end of the first interview, all participating households received a more detailed brochure, giving further information about the survey and thanking respondents for participating.

A minimum of six calls is made at each sampled address before it is considered a non-contact. Interviewers are encouraged to make further calls, if possible. If there is a potential for success, a special conversion letter is sent to households which had refused to participate or had not been contacted. Post interview quality control is carried out with a telephone recall on 10% of all completed interviews.

Interviewers upload their work daily, including information about all the calls they have made, whether or not there was any response. This information is collated by NatCen to construct a weekly field progress monitor report for ISER.

During the second year of Wave 3 (2012) a telephone “mop-up” was introduced. This was started in April, but also covered the sample from January-March. The aim of this was to contact adults who could not be contacted by face-to-face interviewers during the main fieldwork period. Adults in households that were non-responding in the main fieldwork period, except those who had adamantly refused or were deemed to be mentally or physically incapable of participating, were contacted by the NatCen Multi-Mode Unit (MMU) based in Brentwood. The trained and briefed telephone interviewers at the MMU introduce themselves, remind the sample members of the survey, and ask whether they would be able to do the interview by telephone. The purpose of the mop-up” was to increase participation among those who were hard to contact in person.

Analysis by NatCen indicates that the telephone mop-up increased the overall household response rate for that period by about 3 percentage points for the EMB and by just less than 2 percentage points for the GPS. This mop-up was not conducted with the BHPS sample in Wave 3 since they are interview in the first year of each wave.

Towards the end of Wave 3, September 2012, a trial was conducted in two field areas in which an additional incentive was used at the re-issue stage. This was then rolled-out across the sample from October, and so covers the last quarter of Wave 3. In the implementation, non-responding households were reviewed by the NatCen Operations Department in Brentwood for re-issue and possible re-allocation to a different interviewer. Households which had refused to participate in the initial fieldwork period, but where the assessment was that this was a “soft” refusal, were sent a re-issue letter which mentioned an additional incentive if they participated during the re-issue fieldwork period. Other non-responding households were sent a normal re-issue letter, but the interviewers had discretion to offer the additional incentive on the door-step if they felt that this would convert a non-responding household to a participating household.

In addition, during the latter quarter of Wave 3 fieldwork, more effort was made to increase interviewer continuity for households across waves, rather than prioritising interviewer efficiency. It is estimated that these two procedures, which were launched almost simultaneously, increased household response rates by around 4 percentage points for the EMB sample and by around 2.5 percentage points in the GPS in Quarter 8. The procedures adopted in Wave 3 to maintain household response were continued in the Wave 4 fieldwork.

2.3.3. PANEL MEMBERSHIP AND PANEL MAINTENANCE

The rules for following individual respondents over time are based upon the composition of the household. Individuals found at selected households in the first

wave were designated as Original Sample Members (OSM). We attempt to maintain OSM respondents as part of the sample as long as they live in the UK. In addition, births to an OSM mother are classified as OSM. Individuals joining the household of an OSM after enumeration of the household at Wave 1 are Temporary Sample Members (TSM). One deviation from this is for individuals who were not an ethnic minority within the households selected as the EMB sample. At Wave 1, these individuals were classified as TSMs. We attempt to interview TSM participants in successive waves as long as they live in the household of an OSM.

A male TSM who fathers a child with an OSM female becomes a Permanent Sample Member (PSM). PSMs are treated in the same way as OSMs in the following rules. In sum, TSMs are not followed for interviews when they leave the household, but OSMs and PSMs are.

For panel maintenance, ISER maintains a database of information on respondents so we can send communications to them and to allocate interviewers. This information is vital for minimising attrition. The data base builds on contact information collected during the survey interviews, and is updated throughout the year. There are, for example, new addresses, household splits and moves out of the country or into an institution. Change of address cards were also returned to ISER in cases where a whole household moved or a new resident returned the card giving the forwarding address. It is possible for ISER to be notified of some deaths through these means.

A between wave mailing is also used to help maintain contact with participants and update addresses. The mailing has a report of research findings, an address confirmation slip and materials to encourage registration with the participant website. The participant website can be seen at
https://www.understandingsociety.ac.uk/participants

2.3.4. RESPONSE OUTCOMES

2.3.4.1. Wave 1

The Wave 1Main Survey fieldwork started on 8th January 2009 and ended on the 7th March 2011 (including the re-issue period). In total, interviews were achieved in 30,169 households (26,089 in the GPS, 4,080 in the EMB sample), with full or proxy interviews with 50,994 individuals (43,674 in the GPS and 7,320 in the EMB sample).

Table 2 and Table 3 present the household and individual response rates for Wave 1. The individual response rates are for co-operating households only.

Table 2: Household response rates among eligible households, Wave 1

	GPS			EMB
	Great Britain	Northern Ireland	Total	
Responding	57.1%	60.9%	57.3%	39.9%
Non-contact	8.1%	11.0%	8.3%	28.0%
Refusal	33.9%	27.4%	33.6%	29.0%
Other	0.8%	0.7%	0.8%	3.1%
N	43,267	2,107	45,374	10,077

The response rates for the EMB sample component do not make any correction for the probability of non-interviewed cases being ineligible. The estimated response rate taking this factor into account is substantially higher.

Table 3: Individual response rates, Wave 1

	GPS			EMB
	Great Britain	Northern Ireland	Total	
Full interview	82.0%	77.3%	81.8%	72.4%
Proxy interview	5.3%	3.5%	5.2%	6.9%
Refusal	6.5%	9.2%	6.7%	8.7%
Other non-interview	6.1%	9.9%	6.3%	12.1%
N	47,615	2,584	50,199	9,237

2.3.4.2. Wave 2

The Wave 2 main survey fieldwork started on 12th January 2010 and ended on the 27th March 2012 (including the re-issue period).

Household response rates for Wave 2 of the UKHLS are shown in Table 4. The table separates the different samples. The GPS consists of respondents in Great Britain and Northern Ireland. The EMB households are only located in Great Britain. The former-BHPS sample consists of the *Living in Britain* sample (started in 1991), the *Living in Scotland* and *Living in Wales* boost samples (started in 1999) and the *Northern Ireland Household Panel Survey* (NIHPS, started in 2001), also a boost sample.

Ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals or emigrated from the UK. For the former-BHPS component, ineligible households would also include households which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

Fully responding households are those in which the household is successfully enumerated, the household questionnaire is completed, and all eligible adults give an individual interview. Partially responding households are those where the household is enumerated and a household questionnaire is done, and at least one eligible adult but not all eligible adults complete an individual interview.

Household response rates were higher in Northern Ireland than in the rest of the UK. The household response rate for the continuing *Understanding Society* GPS was 76.8% in Great Britain and 81.9% in Northern Ireland. The household response rates for the former-BHPS component were similar to the *Understanding Society* GPS. Among the samples in Great Britain, the *Living in Britain* households had the highest response rate at 77.2%. The *Living in Wales* households had a similar response rate to the *Living in Britain* sample (76.8%), whilst *Living in Scotland* had a lower response rate at 73.5%. The NIHPS had a higher household response rate than in Great Britain, with 84.8%. The response rates for the BHPS samples in Great Britain were disappointing, given that this was, in effect, Wave 19 for many households.

However, the lower response rate may have been due to the change in the fieldwork agency, interviewers, survey name, and logo. Interestingly, in Northern Ireland where the survey name and logo changed, but the fieldwork agency and so the interviewer stayed the same as in NIHPS, the response rate was much higher.

Table 4: Household response rates, Wave 2*

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	16,003 61.8%	873 65.3%	2,030 49.3%	3,112 66.5%	793 64.3%	833 64.1%	990 73.4%	24,634 61.7%
Partially responding	3,888 15.0%	221 16.5%	749 18.2%	504 10.8%	114 9.2%	165 12.7%	153 11.4%	5,794 14.5%
All responding	19,891 76.8%	1,094 81.9%	2,779 67.5%	3,616 77.2%	907 73.5%	998 76.8%	1,143 84.8%	30,428 76.2%
Non-contact	1,116 4.3%	22 1.7%	299 7.3%	217 4.6%	73 5.9%	62 4.8%	33 2.5%	1,825 4.6%
Untraced mover	1,450 5.6%	50 3.7%	411 10.0%	181 3.9%	49 4.0%	50 3.9%	43 3.2%	2,235 5.6%
Refusal	3,359 13.0%	162 12.1%	600 14.6%	648 13.8%	199 16.1%	185 14.2%	117 8.7%	5,281 13.2%
Other non-interview	94 0.4%	8 0.6%	28 0.7%	20 0.4%	6 0.5%	5 0.4%	12 0.9%	173 0.4%
Total*	25,910	1,336	4,117	4,682	1,234	1,300	1,348	39,942

* Base is all households issued to the field for wave 2, minus any found to have become ineligible.

Non-contact rates were lower in Northern Ireland than in Great Britain. The level of untraced movers was higher for the UKHLS GPS in Great Britain than in the former-BHPS. The levels of non-contact and untraced movers were highest in the EMB samples, possibly reflecting their younger average age, concentration in large urban areas, and higher level of mobility. Within the former-BHPS, the level of untraced movers was higher than in the past. This is likely to be due to the longer gap between waves of interview. The interviews for the former BHPS sample for Wave 2 of the UKHLS took place throughout 2010 and into the early months of 2011. The previous interview for most of these households was between September and December 2008. As the gap between the Wave 18 BHPS interview and the Wave 2 *Understanding Society* interview increased, so did the level of untraced movers. Refusals, as well, were generally higher in Great Britain than in Northern Ireland. Refusals are expected to be higher at the second wave of a longitudinal study than at subsequent waves. The higher than expected refusal rate for the former BHPS sample, particularly those in Great Britain, may be due to the aforementioned change in the name and logo of the study as well as the change in fieldwork agency and thus, for most households, a change of interviewer.

Table 5 shows the Wave 2 cross-sectional response rates for adults. Where a household responded, we have an individual-level outcome for all adults. Where a

household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household. From this we can see, for example, that we were not able to interview 7,229 adults in the UKHLS GPS in Great Britain because they were residing in households who refused to participate at Wave 2. In the Great Britain samples of the former BHPS there is a relatively small group of households who only give telephone interviews.

On a longitudinal study, such as the UKHLS, researchers are typically interested in having pairs of observations on the same individual to investigate individual-level change over time. Table 5 takes as the baseline all those who gave a full interview at the previous wave, and shows their outcome at Wave 2. For the former BHPS samples, the previous wave was Wave 18 (i.e., for the *Living in Britain sample*), Wave 10 (i.e., for the *Living in Scotland* and *Living in Wales samples*) or Wave 8 (i.e., for the NIHPS); all collected in 2008. Once more, we see that there is a higher re-interview rate in the Northern Ireland samples than in Great Britain. The lowest re-interview rate is in the EMB sample, largely due to a higher level of non-contacted households or households who moved but could not be traced. Interestingly, the re-interview rate was higher in the General Population GB sample than in the three samples that made up the former-BHPS GB samples. Overall, in the Waves 1 and 2 data, pairs of observations are available for 45,836 adults. If proxy and telephone interviews are included, this increases to 47,282 adults.

For more detail, please see the working paper on nonresponse and attrition (Lynn, Burton et al. 2012)

Table 5: Wave 2 Cross-sectional individual adult response rates by sample origin

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	32,381 60.8%	1,770 62.3%	4,978 46.3%	6,140 61.6%	1,461 58.1%	1,651 59.6%	2,008 71.7%	50,389 59.4%
Proxy interview	2,722 5.1%	87 3.1%	615 5.7%	253 2.5%	49 2.0%	86 3.1%	58 2.1%	3,870 4.6%
Telephone interview	--	--	--	202 2.0%	66 2.6%	58 2.1%	--	326 0.4%
Other non-interview	1,184 2.2%	126 4.4%	472 4.4%	200 2.0%	58 2.3%	64 2.3%	107 3.8%	2,211 2.6%
Refusal	2,104 4.0%	218 7.7%	511 4.8%	341 3.4%	92 3.7%	114 4.1%	133 4.8%	3,513 4.1%
Household non-contact	3,338 6.3%	125 4.4%	1,156 10.8%	555 5.6%	210 8.3%	155 5.6%	111 4.0%	5,650 6.7%
Household refusal	7,229 13.6%	350 12.3%	1,743 16.2%	1,493 15.0%	400 15.9%	427 15.4%	203 7.2%	11,845 14.0%
Household other non-interview	118 0.2%	5 0.2%	60 0.6%	29 0.3%	9 0.4%	6 0.2%	4 0.1%	231 0.3%
Household untraced	4,178 7.9%	159 5.6%	1,207 11.2%	754 7.6%	172 6.8%	208 7.5%	178 6.4%	6,856 8.1%
Total	53,254	2,840	10,742	9,967	2,517	2,769	2,802	84,891

Table 6: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at Wave 1

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	29,646 74.3%	1,640 81.0%	4,200 62.2%	5,633 69.4%	1,335 67.8%	1,507 67.6%	1,875 83.3%	45,836 72.4%
Proxy interview	775 1.9%	16 0.8%	188 2.8%	97 1.2%	17 0.9%	38 1.7%	15 0.7%	1,146 1.8%
Telephone interview	--	--	--	184 2.3%	59 3.0%	57 2.6%	--	300 0.5%
Other non-interview	334 0.8%	22 1.1%	157 2.3%	73 0.9%	16 0.8%	28 1.3%	32 1.4%	662 1.1%
Refusal	316 0.8%	25 1.2%	94 1.4%	53 0.7%	13 0.7%	15 0.7%	11 0.5%	527 0.8%
Household non-contact	1,890 4.7%	68 3.4%	500 7.4%	376 4.6%	126 6.4%	96 4.3%	34 1.5%	3,092 4.9%
Household refusal	4,144 10.4%	167 8.3%	734 10.9%	965 11.9%	245 12.4%	260 11.7%	109 4.8%	6,633 10.5%
Household other non-interview	65 0.2%	2 0.1%	31 0.5%	18 0.2%	3 0.2%	2 0.1%	2 0.1%	123 0.2%
Household untraced	2,252 5.6%	63 3.1%	639 9.5%	439 5.4%	105 5.3%	140 6.3%	105 4.7%	3,744 5.9%
Household ineligible	507 1.3%	22 1.1%	208 3.1%	280 3.5%	51 2.6%	86 3.9%	68 3.0%	1,222 1.9%
Total	39,929	2,025	6,751	8,118	1,970	2,229	2,251	63,285

2.3.4.3. Wave 3

The Wave 3 main survey fieldwork started on 7th January 2011 and ended on the 12th July 2013 (including the re-issue period). Table 7 shows the household response rates for Wave 3 of the UKHLS. The table separates the different samples as in Table 4, above.

As before, ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals, emigrated from the UK or which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

Household response rates (including partial household response) were higher in Northern Ireland than in the rest of the UK, as at Wave 2. The household response rate for the continuing *Understanding Society* GPS was 75.3% in Great Britain, and 79.1% in Northern Ireland. The household response rates for the former-BHPS samples were higher than the *Understanding Society* GPS, both in terms of overall household rates and fully responding households, although there is only a slight difference between the overall household rate for Living in Scotland and the British GPS rate. Among the samples in Great Britain, the *Living in Britain* households – who have been part of the sample for the longest time – had the highest response rate at 81.9%. The *Living in Wales* households had a similar response rate to the *Living in Britain* sample (80.9%), whilst *Living in Scotland* had a lower response rate at 76.5%. The NIHPS had a higher household response rate than the Great Britain samples, with 86%.

Table 7: Household response rates, Wave 3*

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	13,629 57.1%	680 55.3%	1,566 42.9%	2,762 66.2%	664 62.2%	726 62.2%	833 66.4%	20,860 57.3%
Partially responding	4,341 18.2%	293 23.8%	951 26.0%	653 15.7%	153 14.3%	218 18.7%	246 19.6%	6,855 18.8%
All responding	17,970 75.3%	973 79.1%	2,517 68.9%	3,415 81.8%	817 76.5%	944 80.9%	1,079 86.0%	27,715 76.1%
Non-contact	1,056 4.4%	91 7.4%	277 7.6%	186 4.5%	77 7.2%	55 4.7%	54 4.3%	1,796 4.9%
Untraced mover	1,458 6.1%	44 3.6%	351 9.6%	133 3.2%	48 4.5%	50 4.3%	51 4.1%	2,135 5.9%
Refusal	3,088 12.9%	102 8.3%	461 12.6%	395 9.5%	115 10.8%	105 9.0%	55 4.4%	4,321 11.9%
Other non-interview	295 1.2%	19 1.6%	47 1.3%	44 1.1%	11 1.0%	13 1.1%	15 1.2%	444 1.2%
Total*	23,867	1,229	3,653	4,173	1,068	1,167	1,254	36,411

* Base is all households issued to the field for Wave 3, minus any found to have become ineligible.

The EMB had the lowest household response rate of all the samples in the study. The main reasons for the lower response appear to be a higher level of non-contacts and untraced movers in this sample. The EMB sample is concentrated in areas of high ethnic minority density, which tend to be very urbanized areas, particularly in London. Residential mobility is higher among those living in these types of area, which may contribute to the higher levels of untraced movers. Non-contact rates also tend to be higher in cities, particularly in areas of tower blocks, flats or apartments with a common – often locked – entrance. The refusal rates for the EMB are similar to those of the GPS in Britain and only a couple of percentage points higher than the *Living in Scotland* sample.

Refusal rates were still higher for the former-BHPS samples in Britain than they had been in the later years of the BHPS. The refusal rate from Waves 14-18 were around 6%, compared to 9%-11% at Wave 3 of *Understanding Society*. This may reflect some resistance after the change in fieldwork agency.

Table 8 shows the cross-sectional response rates for adults in Wave 3. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household. At Wave 3 the telephone interview was conducted using the same instrument as the face-to-face instruments but with slight changes to reflect the aural rather than visual delivery of the questionnaire (e.g., there were no showcards). Apart from these changes, the content of the telephone questionnaire was the same as the face-to-face questionnaire and so these interviews are classified as “full interviews” below.

In a longitudinal study, such as the UKHLS, researchers are typically interested in having pairs of observations on the same individual to investigate individual-level change over time. Table 9 takes as the baseline all those who gave a full interview at the previous wave, and shows their outcome at Wave 3. Once more, we see that there is a higher re-interview rate in the Northern Ireland samples than in Great Britain. The lowest re-interview rate is in the EMB sample, largely due to a higher level of refusal households or households who moved but could not be traced. Unlike at Wave 2, the re-interview rate was lower in the General Population GB sample than in the three samples that made up the former-BHPS GB samples. The lower re-interview rate at Wave 2 for these BHPS sample households may reflect a “blip” caused by the changes in fieldwork agency, the name and branding of the survey. Overall, in the Waves 2 and 3 data, pairs of observations are available for 45,903 adults. If proxy interviews are included, this increases to 49,708 adults.

Table 8: Wave 3 cross-sectional individual adult response rates by sample origin

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	29,135 61.3%	1,550 59.5%	4,432 47.8%	5,959 70.4%	1,377 66.0%	1,604 67.6%	1,846 72.5%	45,903 61.3%
Proxy interview	2,516 5.3%	47 1.8%	667 7.2%	333 3.9%	65 3.1%	107 4.5%	70 2.8%	3,805 5.1%
Other non-interview	1,039 2.2%	150 5.8%	426 4.6%	144 1.7%	54 2.6%	60 2.5%	107 4.2%	1,980 2.6%
Refusal	2,051 4.3%	207 7.9%	490 5.3%	322 3.8%	68 3.3%	111 4.7%	142 5.6%	3,391 4.5%
Household non-contact	2,274 4.8%	253 9.7%	834 9.0%	418 4.9%	176 8.4%	135 5.7%	134 5.3%	4,224 5.6%
Household refusal	7,826 16.5%	295 11.3%	1,611 17.4%	1,005 11.9%	260 12.4%	255 10.7%	149 5.9%	11,401 15.2%
Household other non-interview	414 0.9%	34 1.3%	107 1.2%	67 0.8%	15 0.7%	22 0.9%	25 1.0%	684 0.9%
Household untraced	2,262 4.8%	71 2.7%	698 7.5%	219 2.6%	71 3.4%	80 3.4%	73 2.9%	3,474 4.6%
Total	47,517	2,607	9,265	8,467	2,086	2,374	2,546	74,862

Table 9: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at Wave 2

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	25,853 78.5%	1,440 80.5%	3,519 69.4%	5,424 83.8%	1,249 80.6%	1,434 82.3%	1,747 85.3%	40,666 78.8%
Proxy interview	706 2.1%	16 0.9%	201 4.0%	99 1.5%	22 1.4%	37 2.1%	34 1.7%	1,115 2.2%
Other non-interview	295 0.9%	48 2.7%	109 2.2%	56 0.9%	14 0.9%	22 1.3%	27 1.3%	571 1.1%
Refusal	287 0.9%	30 1.7%	77 1.5%	47 0.7%	10 0.7%	17 1.0%	15 0.7%	483 1.0%
Household non-contact	832 2.5%	44 2.5%	212 4.2%	158 2.4%	57 3.7%	51 2.9%	34 1.7%	1,388 2.7%
Household refusal	2,919 8.9%	108 6.0%	418 8.2%	312 4.8%	99 6.4%	82 4.7%	59 2.9%	3,997 7.8%
Household other non-interview	258 0.8%	21 1.2%	40 0.8%	42 0.7%	11 0.7%	15 0.9%	22 1.1%	409 0.8%
Household untraced	1,285 3.9%	52 2.9%	352 6.9%	230 3.6%	52 3.4%	54 3.1%	78 3.8%	2,103 4.1%
Household ineligible	483 1.5%	29 1.6%	142 2.8%	108 1.7%	36 2.3%	31 1.8%	33 1.6%	862 1.7%
Total	32,918	1,788	5,070	6,476	1,550	1,743	2,049	51,594

2.3.4.4.Wave 4

The Wave 4 main survey fieldwork started on 8h January 2012 and ended on the 19th June 2013 (including the re-issue period). Table 10 shows the household response rates for Wave 4 of the UKHLS. The table separates the different samples as in Table 7, above.

As before, ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals, emigrated from the UK or which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

The fully-responding household response rates for the former-BHPS samples were higher than the *Understanding Society* samples. The household response rate was lower in the former *Living in Scotland* than the other BHPS samples.

Table 10: Household response rates, Wave 4*

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	12,897 62.0%	603 56.9%	1,505 46.9%	2,579 68.9%	594 63.2%	666 62.7%	789 67.9%	19,633 61.4%
Partially responding	3,958 19.0%	233 22.0%	819 25.5%	600 16.0%	146 15.5%	216 20.3%	209 18.0%	6,181 19.3%
All responding	16,855 81.0%	836 78.9%	2,324 72.4%	3,179 84.9%	740 78.7%	882 83.0%	998 85.9%	25,814 80.7%
Non-contact	858 4.1%	63 5.9%	244 7.6%	113 3.0%	41 4.4%	39 3.7%	45 3.9%	1,403 4.4%
Untraced mover	754 3.6%	50 4.7%	200 6.2%	87 2.3%	33 3.5%	39 3.7%	46 4.0%	1,209 3.8%
Refusal	2,134 10.3%	96 9.1%	392 12.2%	325 8.7%	100 10.6%	94 8.9%	53 4.6%	3,194 10.0%
Other non-interview	202 1.0%	15 1.4%	50 1.6%	38 1.0%	26 2.8%	8 0.8%	20 1.7%	359 1.1%
Total*	20,803	1,060	3,210	3,742	940	1,062	1,162	31,979

* Base is all households issued to the field for Wave 4, minus any found to have become ineligible.

The EMB continues to have the lowest household response rate of all the samples in the study. The main reason for the lower response continues to be a higher level of non-contacts and untraced movers in this sample. The refusal rates for the EMB at Wave 4 are slightly higher than those of the GPS in Britain and only a couple of percentage points higher than the *Living in Scotland* sample.

Table 11 shows the cross-sectional response rates for adults in Wave 4. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household.

Table 11: Wave 4 Cross-sectional individual adult response rates by sample origin

	UKHLS GP sample		EMB	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	27,643 67.0%	1,341 60.7%	4,236 52.0%	5,547 73.4%	1,222 66.9%	1,479 68.7%	1,749 74.1%	43,217 66.0%
Proxy interview	2,654 6.4%	36 1.6%	666 8.2%	330 4.4%	74 4.1%	115 5.3%	40 1.7%	3,915 6.0%
Other non-interview	740 1.8%	113 5.1%	290 3.6%	104 1.4%	53 2.9%	52 2.4%	88 3.7%	1,439 2.2%
Refusal	1,676 4.1%	180 8.2%	394 4.8%	298 3.9%	62 3.4%	115 5.3%	153 6.5%	2,878 4.4%
Household non-contact	1,958 4.8%	150 6.8%	756 9.3%	243 3.2%	84 4.6%	95 4.4%	110 4.7%	3,396 5.2%
Household refusal	5,209 12.6%	281 12.7%	1,328 16.3%	842 11.1%	243 13.3%	224 10.4%	139 5.9%	8,266 12.6%
Household other non-interview	296 0.7%	35 1.6%	127 1.6%	52 0.7%	40 2.2%	17 0.8%	27 1.1%	594 0.9%
Household untraced	1,076 2.6%	72 3.3%	346 4.3%	143 1.9%	50 2.7%	56 2.6%	56 2.4%	1,799 2.8%
Total	41,252	2,208	8,143	7,559	1,828	2,153	2,362	65,505

Table 12 shows the individual re-interview rates, by sample. The re-interview rate at Wave 4 is higher than it was at Wave 3; 82.7% of those who gave a full interview at Wave 3 also gave one at Wave 4, compared to a 78.8% re-interview rate between Waves 2 and 3. In the UKHLS GPS, the re-interview rate is higher for the British sample than the Northern Ireland sample. The opposite is the case for the former-BHPS samples; the Northern Ireland sample has the highest re-interview rate.

Table 12: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at Wave 3

	UKHLS GP sample		**EMB**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	24,705 83.7%	1,246 79.5%	3,375 74.6%	5,070 84.3%	1,095 78.7%	1,351 82.2%	1,641 87.5%	38,483 82.7%
Proxy interview	609 2.1%	10 0.6%	137 3.0%	101 1.7%	26 1.9%	30 1.8%	7 0.4%	920 2.0%
Other non-interview	198 0.7%	25 1.6%	86 1.9%	30 0.5%	18 1.3%	11 0.7%	24 1.3%	392 0.8%
Refusal	183 0.6%	18 1.2%	62 1.4%	29 0.5%	12 0.9%	8 0.5%	12 0.6%	324 0.7%
Household non-contact	663 2.3%	60 3.8%	201 4.4%	79 1.3%	25 1.8%	26 1.6%	41 2.2%	1,095 2.4%
Household refusal	1,838 6.2%	131 8.4%	384 8.5%	385 6.4%	116 8.3%	106 6.5%	57 3.0%	3,017 6.5%
Household other non-interview	164 0.6%	15 1.0%	57 1.3%	32 0.5%	30 2.2%	7 0.4%	21 1.1%	326 0.7%
Household untraced	857 2.9%	40 2.6%	167 3.7%	221 3.7%	50 3.6%	88 5.4%	41 2.2%	1,464 3.1%
Household ineligible	315 1.1%	23 1.5%	58 1.3%	69 1.2%	19 1.4%	17 1.0%	32 1.7%	533 1.1%
Total	29,532	1,568	6,016	1,391	1,644	1,876	4,527	46,554

2.4. DATA PROCESSING AND CLEANING

The data for a sample month is delivered to ISER in batches, scheduled for 4 months following the beginning of the fieldwork process. This interval allows time for interview re-issue, coding, and data entry from paper documents, e.g., the self-completion instruments. Data is delivered as SPSS system files, which are then exported to triple-S data exchange format and imported into a SIR database.

Quality control processes include extensive data checking to ensure that the data conform to the expected structure and to the routing and range constraints defined by the questionnaire specifications. Data anomalies are investigated to determine whether they are related to:

- the invalid specification of the questionnaire;
- the incorrect scripting of the questionnaire;
- a failure to specify that a particular constraint should be included in the questionnaire;
- an incorrect implementation of the check, or;
- a problem in exporting and/or delivering the data.

After investigation, steps may include correcting the specification, data editing, reporting the error to NatCen to be fixed in a subsequent delivery and/or a quality feedback report suggesting changes to the questionnaire or field practice in subsequent waves.

Batch-specific databases are merged into a single database, from which anonymised data is exported for the creation of public use files. Data distributions are also checked for theoretical and statistical plausibility. This checking is done through direct scrutiny and by analyses which “road-test” the data.

Last but not least, data are being routinely checked in the process of creating added-value content, such as the respondent’s age and sex, based on the information collected across all waves, or pointers to specific other members in the household such as a biological parent. See Section 3.2.4.

2.4.1. CODING

Understanding Society collects freetext information on respondents' job titles and the industry of the job held. Industry descriptions are coded to ONS Standard Industry Code 2007, or SIC 2007. Job titles are coded to the ONS Standard Occupational Classification 2000, or SOC 2000. Coding is undertaken using the Computer Assisted Structured Coding Tool (CASCOT) system.

Several questions, e.g. country of birth, religion, political party, national identity, and citizenship had an “other, please specify” option. These responses were coded using an automated process.

Coding was also done for an open-ended question which read “We've asked you a lot of questions but we also want to know what has happened in your own life that has been especially important to you. Can you please tell me anything that has happened to you, or your family, over the past year that has stood out as important?” The respondent could give up to four answers. The answers were recorded verbatim and manually coded for type of event and its subject.

2.5. DOCUMENTATION OF THE SURVEY INSTRUMENTS

The text of the questionnaires in PDF format is part of the documentation provided through the UK Data Service. Questionnaires can also be found: https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaires. There are household and individual questionnaires and the adult and youth self-completion instruments. The instruments are an important source of information about the wording of individual questions, who was asked, and what questions precede and follow.

Most of the interview is conducted with a computer-assisted personal interview (CAPI). The CAPI instrument governs the flow of questions and recording of answers, but it is not convenient for documentation. On the study website, we present the questionnaire in PDF format. Similar to other PDF documents, the text of the questionnaire can be searched for specific words, such as variable names or words in questions. The PDF self-completion instruments correspond to the way they appeared to participants (except they have been annotated with variable names).

The principal adult questionnaires are organized in modules. Modules can be searched for in the online documentation system https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation.

Instruments and survey materials were translated into multiple languages: Bengali and Punjabi in Urdu and Gurmukhi scripts; Welsh, Arabic, Somali, Cantonese, Urdu and Gujarati. Translated documents can be requested by email from info@understandingsociety.ac.uk.

2.5.1. READING THE QUESTIONNAIRES

Figure 1: Mark-up of household questionnaire shows a marked up sample page, providing information for how to interpret the questionnaire text. Note that the variable names in the questionnaire do not have the wave prefix (a_), b_, etc. or in the general form *w_*.

Figure 2: Mark-up of question with looping from individual questionnaire shows a marked-up sample page from the individual interview. The question is more complex. The question is asked about each natural or biological child, so multiple variables are associated with the question for each natural child. The variables are located in the data file ***a_natchild***, which has one record for each natural child.

Figure 1: Mark-up of household questionnaire

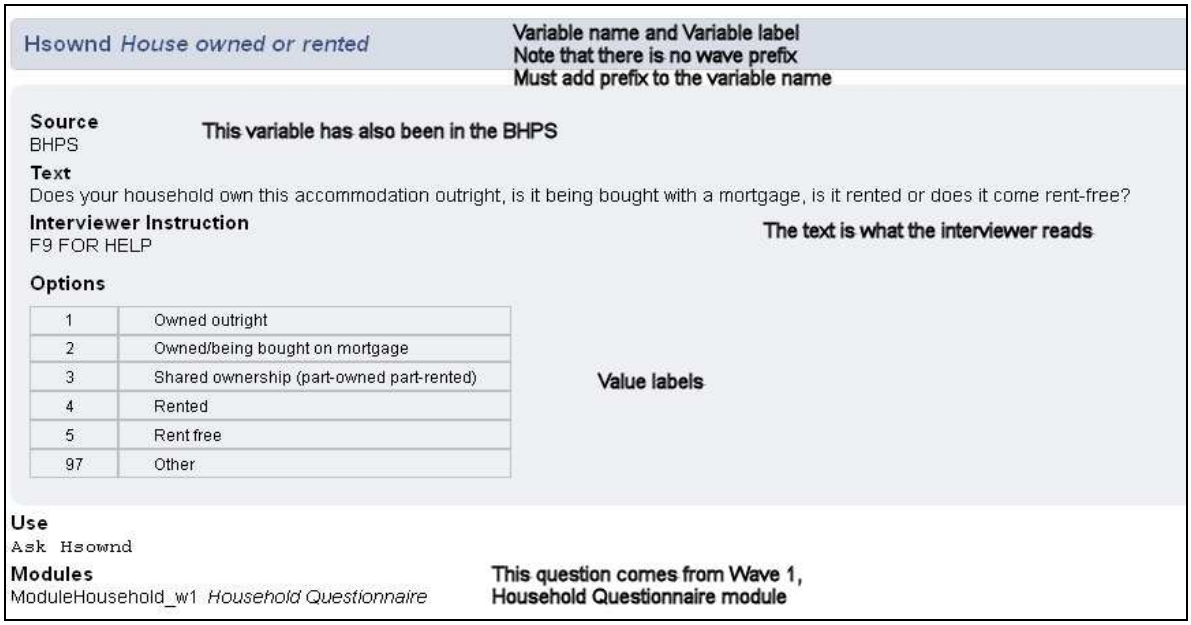
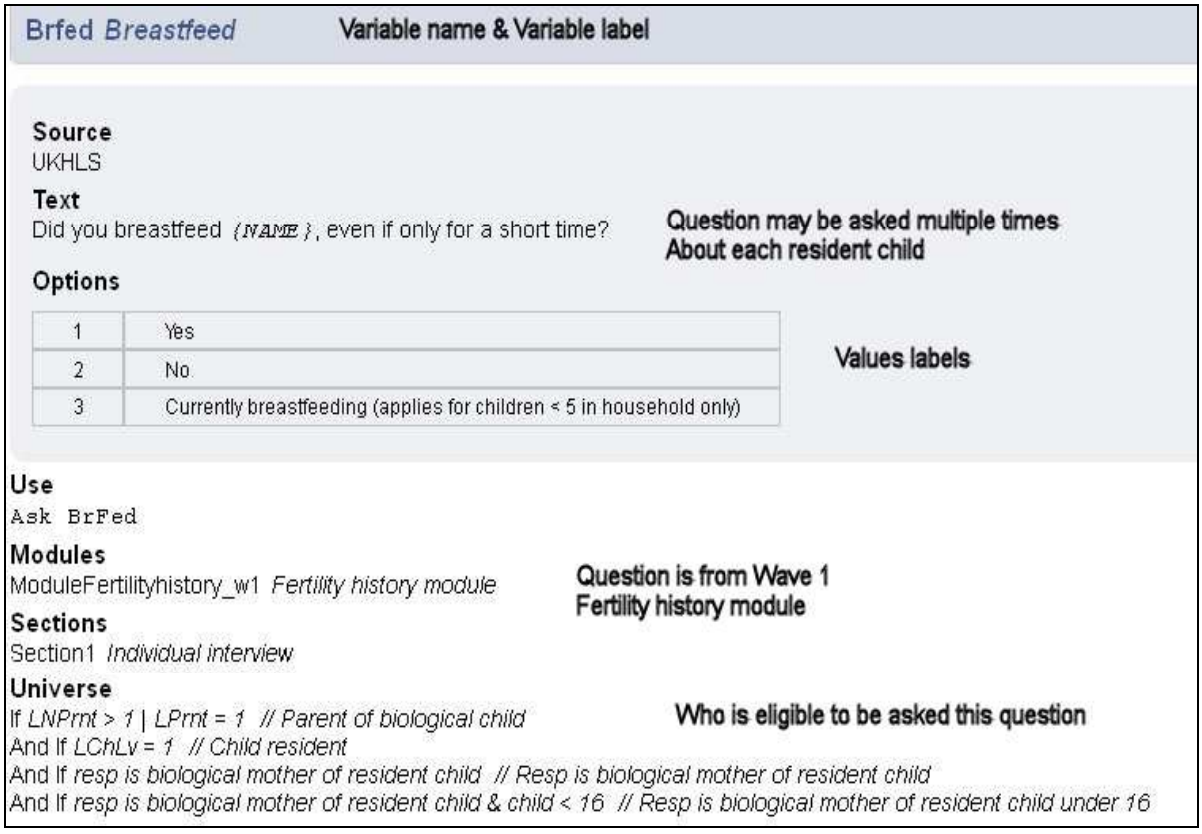

Hsownd House owned or rented
Variable name and Variable label
Note that there is no wave prefix
Must add prefix to the variable name
Source
BHPS
This variable has also been in the BHPS
Text
Does your household own this accommodation outright, is it being bought with a mortgage, is it rented or does it come rent-free?
Interviewer Instruction
F9 FOR HELP
The text is what the interviewer reads
Options
1 Owned outright
2 Owned/being bought on mortgage
3 Shared ownership (part-owned part-rented)
4 Rented
5 Rent free
97 Other
Value labels
Use
Ask Hsownd
Modules
ModuleHousehold_w1 Household Questionnaire
This question comes from Wave 1,
Household Questionnaire module

Figure 2: Mark-up of question with looping from individual questionnaire


Brfed Breastfeed
Variable name & Variable label
Source
UKHLS
Text
Did you breastfeed (NAME), even if only for a short time?
Question may be asked multiple times
About each resident child
Options
1 Yes
2 No
3 Currently breastfeeding (applies for children < 5 in household only)
Values labels
Use
Ask BrFed
Modules
ModuleFertilityhistory_w1 Fertility history module
Question is from Wave 1
Fertility history module
Sections
Section1 Individual interview
Universe
If LNPrnt > 1 | LPrnt = 1 // Parent of biological child
Who is eligible to be asked this question
And If LChLv = 1 // Child resident
And If resp is biological mother of resident child // Resp is biological mother of resident child
And If resp is biological mother of resident child & child < 16 // Resp is biological mother of resident child under 16

2.5.2. SUMMARY OF QUESTIONNAIRE MODULES

About half of the questionnaire content is collected annually, with additional modules collected at different intervals, often every two to three years. The long-term content plan summarizes the pattern that has been collected or planned: https://www.understandingsociety.ac.uk/system/uploads/assets/000/000/018/original/Long_Term_Content_Plan_Nov2011-3.pdf?1355920157.

Table 13 shows that some modules are annual, such as Current Employment. We start to see the repetition of rotating modules on alternating waves, for example, Family Networks, Parents and Children, and Harassment.

Table 13: Summary of questionnaire modules

Module	Wave 1	Wave 2	Wave 3	Wave 4
Demographics	X	X – NE	X - NE	X
Initial Conditions	X	X – NE	X - NE	X – NE
Family background	X		X – NE or not interviewed	X – NE or not interviewed
Ethnicity and National Identity	X	X – NE	X – NE	X – NE
Language	X	see Childhood language		
Religion	X	X - NE in EMB, LDA, GPC or NI	X - NE in EMB, LDA, GPC or NI	X
Religious Practice				X – EMB, LDA, GPC, recent immigrant
Migration History	X			
Partnership History	X	X- NE	X- NE	X- NE
Fertility History	X	X- NE	X- NE	X- NE
Health and Disability	X	see General Health module, Disability module	see General Health module, Disability module	
Caring	X	X	X	X
Employment Status History	X (asked of first 6 months of first year)		see Own First Job module	see Own First Job module
Current Employment	X	X	X	X
Employees	X	X	X	X
Self-employment	X	X	X	X
Job Satisfaction	X	X	X	X

Non-employment	X	X	X	X
Second jobs	X	X	X	X
Discrimination	X - EMB, GPC, LDA		X - EMB, GPC, LDA	
Childcare	X	X	X	X
Parents and Children	X, responsible mother and responsible father of children		X, responsible mother and responsible father of children	
Family Networks	X		X	
Remittances	X			X
Unearned Income and State Benefits	X	X	X	X
Household Finances	X	X	X	X
Politics	X	X	X	X
Harassment	X, EMB, GPC, LDA		X, EMB, GPC, LDA	
Environmental Behaviour	X			X
Consents for linkage to health administrative records	X			X
Consents for linkage to education administrative records	X			X
Consents for linkage to benefit administrative records				X
Interviewer observations	X	X	X	
Proxy	X	X	X	
Childhood language		X		
Ethnic identity		X – EMB, GPC, LDA		
General Health		X	X (see Health and Disability module)	X

Nutrition		X		
Physical Activity		X		
Smoking History		X		
Disability		X	X	X
Annual Event History		X – interviewed before	X – interviewed before	X – interviewed before
Commuting Behaviour		X		X
Physical Work		X		
Work Conditions		X		X
Voluntary Work		X		X
Charitable Giving		X		X
Personal Pensions		X		X
Savings		X		X
Retirement Planning		X –age 45/50/55/60/65 and not retired	X –age 45/50/55/60/65 and not retired	X –age 45/50/55/60/65 and not retired
Domestic Division of Labour	One item on hours of housework asked in first 6 months at w1	X		X
Political Engagement		X – EMB, GPC, LDA		
General Election		X – May – Dec. 2010		
Leisure, Culture and Sports		X		
Leisure Access		X		
Positive and Negative Events		X		
Britishness			X – EMB, GPC, LDA	
Own First Job			X	X
Educational Aspirations			X	X
Young Adults			X – age 16-21	X – age 16-21
Local Neighbourhoods			X	
Urban-rural			X	
Social Networks			X	
Groups and Organisations			X	
Health			X – NE	X – NE

Conditions				
Cognitive Ability			X	
Family Access			X	
Child Maintenance			X – biological parent of child/other biological parent not in HH	
Migration Intention			X – age 45/50/60/65 and not retired	
Political Self Efficacy			X	
News and Media Use			X	X
Service Use				X – EMB, LDA, GPC, recent immigrant
Sleep				X
Twin				X – Biological mother of siblings with same birthday
Mother's return to work				X
Transport behaviour				X
Wealth, assets and debt				X
Olympics				X

NE – new entrant
EMB – Ethnic Minority Boost
GPC – General Population Comparison
LDA – Low density ethnic minority area

The paper self-completion questionnaires carried at waves 1 and 2 were not divided into modules. From Wave 3 onwards, the self-completion content was carried as CASI modules, where the participant would answer the questions using the lap-top. Table 14 summarizes the content in waves 1 to 4.

Table 14: Summary of adult self-completion content

	Wave 1	Wave 2	Wave 3	Wave 4
GHQ-12	X	X	X	X
Sleep	X			
Environmental attitudes and beliefs	X			X
Neighbourhood belonging and participation	X		X	
Trust	X			
Life Satisfaction	X	X	X	X
Short Warwick-Edinburgh Mental Well Being Scale	X			X
Attitudes to risk	X			
Partnership relationship: quality, activities, happiness	X		X	
SF-12	in adult interview Health and Disability module	X	X	X
Identity		X		
Alcohol consumption		X		
Feelings of control		X		
Social support		X		
Gender role opinions		X		X
Big 5 Personality			X	
Sex Orientation			X	
Identity			X – age 16-21	
Family Support and Activities			X – age 16-21	
Bullying			X – age 16-21	
Family Relationships			X – age 16-21	

Smoking Behaviour			X – age 16-21	
Alcohol related Behaviour			X – age 16-21	
Drug Use			X – age 16-21	
Social Networks – 3 Best Friends			X	
Non Co-resident Relationship			X	X
Child Development			X – about child age 3, 5 or 8	X – about child age 3, 5 or 8
Parenting Style			X – refers to child age 10	X – refers to child age 10
Work conditions				X
Young adults				X – age 16-21

The content of the youth self-completion instruments is summarized in the long-term content plan: https://www.understandingsociety.ac.uk/system/uploads/assets/000/000/018/original/Long_Term_Content_Plan_Nov2011-3.pdf?1355920157.

A number of articles in the *Understanding Society* Findings publications focus on measures available see https://www.understandingsociety.ac.uk/research/publications/findings/earlyand https://www.understandingsociety.ac.uk/research/publications/findings/2012.

2.5.3. CONTENT HIGHLIGHTS BY WAVE

2.5.3.1. Wave 1

Wave 1 collected important baseline data. Some Wave 1 measures are stable, that is, not time variant. In subsequent waves, we try to collect this type of information from individuals who are new entrants to the study (see modules marked NE in Table 14). Notice that some modules are covered annually beginning in Wave 1. These represent the strongest areas for examining annual change. See for example Disability, Caring, employment related information, childcare, politics and income and benefits. These have been the focus of major longitudinal research in the BHPS and should also be a prominent focus with *Understanding Society*.

Wave 1 also saw the first rotating modules. These included Parents and Children, Family Networks, and Environmental Behaviour. Within the ethnicity strand – that is within the “Extra 5 Minutes”-questions, there are modules about Remittances, Harassment, and Discrimination. The Parents and Children module has content about attitudes and behaviours related to education, activities and interaction with children, and parenting practices.

Some interesting content in the Wave 1 self-completion questionnaire includes measures of sleep behaviour and sleep quality and a subset of items characterizing relationships with partners from the Dyadic Adjustment scale (Spanier 1976). There are also measures of generalised trust and attitudes to risk.

2.5.3.2. Wave 2

In Wave 2 and subsequent waves, there is the Annual Event History module which is asked of persons previously interviewed. It asks about changing circumstances related to moves, marital status or cohabitation, new children including childbirth and pregnancy, new health conditions, educational experiences, and employment changes.

Wave 2 saw the introduction of a set of rotating modules related to health behaviours (nutrition, smoking, physical activity). There are modules about voluntary work and charitable giving and important modules about savings and personal pensions. Retirement planning is an age-triggered module that is taken up again in Wave 3. Within the self-completion questionnaire, there is content on alcohol consumption, dimensions of identity and gender role attitudes.

The EMB and related samples have modules on political engagement and ethnic identity. Wave 2 also has a major module on work conditions that encompasses such topics as payment mechanisms, unions, pensions, work times, autonomy and security, and work stress.

2.5.3.3. Wave 3

There are multiple new modules for Wave 3 including those on local neighbourhoods, content on social networks in the main survey and the self-completion questionnaires, groups and organizations, use of news and media, and political self-efficacy. There is a major module on cognitive ability, see McFall (2013) for more detail about the concepts and measures of cognitive ability.

Important data related to family ties can be found in the parents and children and family networks modules, both of which are repeated from Wave 1. There is also data about child maintenance payments and relationships with children who do not live in the household. The self-completion modules have parents' reports on children including a version of the Strengths and Difficulties questionnaire, and parenting styles. The self-completion modules also include a Big 5 personality measure, sexual orientation, and several modules of questions for young adults which bring questions from the youth questionnaire into the 16-21 age group.

The EMB and related samples have repeat modules on discrimination and harassment and a new module on Britishness.

2.5.3.4. Wave 4

Most modules for Wave 4 have appeared in earlier waves. They include major modules on work conditions (covering, e.g., transport behaviour and job satisfaction) and modules on environmental behaviours and voluntary work. Wave 4 carries for the first time rotating modules on wealth and assets, financial attitudes and behaviours, and credit and debt. Another highlight of the Wave 4 questionnaire is a one-off module on leisure participation focussing on the Olympics 2012, which were held in and around London 27th July - 12th August 2012. The Wave 4 self-completion instrument for adults includes a focus on mental health and well-being and gender role attitudes.

The EMB and related samples have repeat modules on remittances and on ethnic identity and Britishness.

2.5.4. CHANGES TO THE QUESTIONNAIRE

All survey instruments are carefully tested in a pilot so that any issues with question wording and routing, interview flow and timings can be identified and fixed before the survey is rolled out to the main sample, see Section 2.3.2.2. The survey instruments for a wave of data collection are then fixed for the entire fieldwork period and any question (including its routing) is repeated each wave of data collection, as applicable. However, under certain circumstances changes to the questionnaire may be undertaken.

2.5.4.1. Within a Wave

At the end of the first six months of data collection in Wave 1, multiple variables were dropped because of the length of the interview, e.g., cutting of the employment history module. At the same time other modifications were made, e.g., in question format. Notes about these changes have been documented in the variable view of the online documentation system. Also see Section 3.2, below.

2.5.4.2. Across Waves

There sometimes are changes to the questionnaire across waves. This is the case, for instance, when routing errors only became known after data collection had been completed (e.g., in Wave 1 only the proxy interview included the question for whether or not a respondent had access to a car (***w*_drive**) and from Wave 2 onward this information is available for adult and proxy respondents). Another example is the SF 12 which was asked in the main interview with adults in Wave 1 but was shifted to the adult self-completion in Wave 2. Notes about these changes are documented in the variable view of the online documentation system. Also see Section 3.2, below.

The switch in mode from paper self-completion to the CASI on the lap-top in Wave 4 meant that for some questions the response options were presented differently between waves. For example, where response options were arrayed horizontally in the paper self-completion (e.g., satisfaction questions), they were presented vertically in CASI. There is some evidence that this change in the way in which the response options were presented may affect how some people respond to the question (Burton and (Ed.) 2012).

2.5.5. OTHER FIELDWORK MATERIALS

Other fieldwork materials are also on the website: https://www.understandingsociety.ac.uk/documentation/mainstage/fieldwork-documents.

One example is Show cards, which are used to help respondents with their answers. Show cards are referenced in the questionnaire. Project Instructions were prepared for interviewer training and to serve as a resource in data collection. Documents for communicating with participants are also included on this portion of the website. In Wave 1, we asked for consent to link to administrative health and education records. The information leaflets and consent forms are in this section of the study website.

The Address Record Form (ARF) is an important document for recording information about responding and non-responding households. There are many different versions in Wave 1. Interviewers record the call record, observations on characteristics of accommodation and households, and household outcomes on the

ARF. In Wave 1 there were several different versions of the ARF. The first distinction is between the GPS and the EMB sample. The versions labelled ARF EB, for the EMB sample, are longer because they include questions for screening household members for eligibility. ARF's labelled 2 or 3 are for addresses with multiple households and/or dwelling units. Finally, there are versions for ARF EB1 Year 1 or Year 2. This change in form was required by the change in selection criteria implemented in Year 2 of Wave 1, see Berthoud, Fumagalli et al. (2009) for more detail. The ARF screening card was a show card used during the screening interviews. Additional information about completion of the ARF can be found in the Project Instructions for Interviewers.

3. *UNDERSTANDING SOCIETY* DATA

3.1. INFORMATION ABOUT DATA FILES

The data release consists of multiple files in SPSS or Stata formats distributed by the UK Data Service. Data for different waves are released in separate files. File names begin with a prefix designating the wave of data collection ("a_" for the first wave, "b_" for the second wave; in this user guide we have used "*w*_" to denote waves in general). Data collected from different sources (i.e., e.g., the household interview, the adult interview, the youth interview) are stored in separate files. The root filename is fixed over time. For example, individual level data collected from interviews with responding adults in Wave 1 (both years: 2009 and 2010) is stored in the file ***a_indresp*** and individual level data collected from interviews with responding adults in Wave 2 (both years: 2010 and 2011) is stored in the file ***b_indresp***.

Table 15 lists the data files that contain substantive information collected in interviews with responding households and individuals. They are the most likely data files analysts will want to access.

Table 15: List of select data files: Data from responding sample members

Filename	Description
*w*_hhresp	Substantive data from responding households
*w*_indresp	Substantive data for responding adults (16+) including proxies and telephone interviews from individual questionnaires including self-completion
*w*_youth	Substantive data from youth questionnaire (age 10-15)

Table 16 lists some additional data files analysts may need to access. In particular, we would like to point users to the data file **xwavedat**, in Table 17, which contains stable characteristics of individuals, such as ethnicity, which is typically collected only once in the lifetime of the study.

The complete list of files and their descriptors can be seen in the online documentation system: https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation.

Table 16: List of select data files: Data from enumerated sample members

Filename	Description
*w*_indall	Household grid data for all persons in household, including children and non-respondents
*w*_child	Childcare, consents and school information of all children in the household
*w*_egoalt	Kin and other relationships between pairs of individuals in the household

Table 17: List of select data files: Cross-wave files

Filename	Description
xwavedat	Stable characteristics of individuals
xivdata	Interviewer characteristics
xwaveid	Individual and household identifiers across all waves

3.1.1. PARADATA

Some paradata, i.e., additional data collected about the interview process is available. These consist of call records, timings data and other information collected by the interviewers during the interview. The ***w*_callrec** data file has information on the number of calls made as well as the issue number, time and date and the outcome of each call. Information on the date of receipt of the case and the interviewer associated with each issue as well as the outcome at the end of each issue period is available in the file ***w*_issue**. In addition to this, information collected in the address response form (ARF) by interviewers while contacting each household and asking household members to participate in the survey is available in ***w*_hhsamp**. This includes data on the area surrounding the address, the type of accommodation and other information that the interviewer can observe about sampled addresses. Reasons for refusal are also available. Interviewers also collect some information about the quality of the interview and persons present during the interview process. This is available along with substantive data collected during adult individual interviews (including proxy interviews) in ***w*_indresp**.

Table 18: List of select data files: Paradata

Filename	Description
*w*_hhsamp	Data from Address Record File for issued households
*w*_callrec	Information about interview outcome at each call
*w*_issue	Information about interview outcomes at each issue including interviewer number

3.2. INFORMATION ABOUT VARIABLES

3.2.1. LEARNING ABOUT THE STUDY VARIABLES

There are multiple resources for learning about the study variables in order to plan analyses. These include the questionnaires and the module and variable views in the online documentation system.

Many of the basic (non-derived) variables can be learned about directly from the questionnaires. As was shown in Figure 2, the questionnaire has much useful information. Please note that in the questionnaire, the variable name does not have the wave prefix. It also shows the brief variable label, text of the question, source of the question and value labels. Show-cards to help the respondent in answering are also marked as part of the questionnaire. You can go back and forth from the question view to the variable view.

3.2.2. VARIABLE NAMING AND LABELLING CONVENTIONS

Most variables have a mnemonic name. Variables begin with a prefix designating the wave of data collection ("a_" for the first wave, "b_" for the second wave; in this user guide we have used "*w*_" to denote waves in general). To ease identification of groups of variables a number of additional general naming conventions have been applied. For instance, following the wave prefix, information from the self-completion interview with adults starts with the prefix "sc"; information from the interview with young adults with the prefix "ya", and information from the child development module with the prefix "cd". Similarly, we have attempted to include in the variable name the acronym of well-known instruments such as the Strengths and Difficulties Questionnaire (SDQ) or the General Health Questionnaire (GHQ). See, for example, **c_ypsqda** to **c_ypsdqy** on data file **c_youth** or **d_scghq2_dv** on data file **d_indresp**. Most added-value variables, i.e. variables that are produced post-field, are clearly marked in the data by suffixes: weights are shown by the suffixes "_lw" or "_xw"; most derived variables are shown by the suffix "_dv", and pointers to other members in the household typically end on "pno" or "pid". The prefix "ff_" following the wave prefix shows variables that were fed forward from previous waves to route respondents appropriately in the script.

We have attempted to keep the names of variables that came from the BHPS the same, for the convenience of analysts, but this has not always been possible given overruling naming conventions in the UKHLS. To identify corresponding variables in the BHPS, analysts should consult the BHPS documentation: https://www.iser.essex.ac.uk/bhps/documentation/volb/index.html.

Note that the variable name does not change over time so long as the underlying question does not change substantially. Analysts are advised, however, to carefully read the variable notes in the online documentation to keep track of any definitional changes or changes in the code frame that may impact study results. An example is the derived variable ***w*_hiqual_dv** which from Wave 2 onward includes information from the BHPS, and where the code frames on the BHPS and *Understanding Society* do not perfectly align.

3.2.3. VARIABLE VALUES AND LABELS

The detailed variable view provides information about valid and invalid responses. Additional codes denote different types of reasons for the lack of a valid response. These values have not been specified as missing in Stata or SPSS. However, these statistical packages have commands to assign values to missing for many variables simultaneously. Table 19 describes the missing value codes.

Table 19: Missing value codes

Value	Description
-9	Missing by error or implausible
-8	Not applicable to the person or because of routing
-7	Proxy respondent. The question was not asked of proxy respondents or derived variable cannot be computed for proxy respondents.
-2	Refused
-1	Don't know

The meaning of other values is explained with the variable's value labels. There may also be notes in the detailed variable view of the online documentation system on the website: https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation.

Note that the default missing value code for derived variables tends to be -9 "missing or wild". Missing value codes on the youth self-completion questionnaire also tend to be less accurate because the instrument was administered as a paper-and-pencil questionnaire and processing was therefore not as closely monitored, e.g., respondents may not have followed the question routing. We recommend that users carefully read the questionnaires and compare missing value distributions across waves before using the substantive information contained in them.

3.2.4. IDENTIFIERS AND OTHER USEFUL VARIABLES

Households are identified by ***w*_hidp**, a wave specific variable with a different prefix for each wave. It can be used to link information about a household from different records within a wave, but cannot be used to link information across waves. Since the composition of households can change between waves, the data do not include a longitudinal household identifier.

Individuals are identified by the personal identifier (**pidp**), which is consistent in all waves and can be used to link information about a person from different records belonging to one wave, or to link information from different waves. Individuals are

also identified by ***w*_pno** – the person number within the household. The combination of ***w*_hidp** and ***w*_pno** is unique for each individual.

Table 20 lists some variables commonly used in analysis and may help the analyst to begin planning. Recall that the variables with the prefix "*w*_" have the values for that wave. There is also the file **xwavedat** which has variables with stable values. Variables in that file do not have the wave prefix. Analysts should also remember to consult the section on specifying the complex sampling variables from Section 3.6 and on weighting from Section 3.7.

Table 20: Some useful variables

Variable name	Description
pidp	Cross-wave person identifier
pid	Cross-wave person identifier for continuing BHPS sample members
*w*_hidp	Household identifier
*w*_pno	Within wave person number in the household
mpid, fpid	Cross wave identifier of natural mother, father
*w*_ppid, *w*_sppid	Cross-wave identifier of current cohabitee, spouse
*w*_ivfio	Individual response outcome
*w*_psu	Primary sampling unit
*w*_strata	Sampling strata
*w*_hhorig	Sample origin
*w*_psnenus_xw	Weight for household grid and household interview
*w*_indinus_xw	Weight for individual interview
*w*_xtr5min_dv	Identifier for the "Extra 5 minutes" analytical sample
*w*_country	Country of the UK
*w*_gor_dv	Government Office Region
*w*_hhsize	Household size
*w*_hhtype_dv	Household type
*w*_tenure_dv	Housing tenure
*w*_sex	Respondent sex, also see ***w*_sex_cr**
*w*_dvage	Respondent age, also see ***w*_age_cr**
*w*_marstat	Legal marital status
*w*_marstat_dv	De facto marital status
*w*_ukborn	Born in the UK and UK country of birth
*w*_racel_dv	Ethnicity
*w*_jbstat	Current economic activity (employment status)
*w*_jbhas	Did paid work last week
*w*_fihhmngrs_dv	Gross household income in past 30 days
*w*_jbnssec8_dv	Social class (NS-SEC). 8-category version.
*w*_jbsoc00	Current occupation (SOC2000)
*w*_fenow	Still in further education
*w*_qfhigh	Highest educational qualification

*w*_health	Long standing illness or impairment
*w*_scsf12mcs_dv	SF12 Mental health component score
*w*_scsf12pcs_dv	SF12 Physical health component score
*w*_nchild_dv	Number of respondent's natural children in household

3.3. NOTES ON USING THE BHPS SAMPLE

The continuing sample from the British Household Panel Survey (BHPS) joined the *Understanding Society* sample in Wave 2.

Both samples can be used for cross-sectional and longitudinal analyses. For the appropriate weights, select from Table 24 to Table 31.

The cases in the two samples can be distinguished using the variable ***w*_hhorig**. The variable also allows the identification of different components of the BHPS sample (see below).

The questionnaires used for the two samples are the same. There are, however, a few differences in the data collected. One important issue is that the date of previous interview for GPS sample members who were interviewed at the previous was approximately 12 months earlier, while for the former BHPS sample the gap was between 13 and 27 months for sample members interviewed at Wave 18 of BHPS. This means that the reference period for the history of events since the last interview will be longer for the BHPS sample. This variation in the reference period applies only to Wave 2.

For longitudinal analysis of the GPS sample, cases may be matched to Wave 1 data, available as part of this release from the UK Data Service, using the variable **pidp**, the *Understanding Society* cross-wave person identifier.

However for the BHPS sample, a different identifier will need to be used: the variable **pid**, which is the BHPS cross-wave person identifier. The **pid** identifier is available in all person level files in the *Understanding Society* beginning in the Wave 2 release and in the 18-wave BHPS longitudinal data set.

While the great majority of BHPS sample cases who were interviewed in *Understanding Society* Wave 2 were previously interviewed at Wave 18 in 2008-9, there are a number who were last interviewed at an earlier wave. Information about the response status of BHPS sample members at each of the 18 waves is contained in the BHPS file **xwaveid**.

The BHPS data set also contains a file called **xwavedat**, which contains the values for stable variables (e.g. ethnic group, parent social class etc.). Because of some differences in variable definition, this information has not been copied across to the new *Understanding Society* file, also called **xwavedat**. However in most cases values of these variables can be obtained by matching to the BHPS file. We hope to produce a harmonized version at subsequent release.

In matching to earlier waves of BHPS data it is important to be aware that variable names in the BHPS data set have slightly different formats:

- they are limited to eight characters
- there is no underscore separating the wave prefix from the main part of the name
- derived variables, imputation flags, weights and other special variables are not distinguished by _dv or _if suffixes.

However, most questionnaire variables which are carried in both surveys will have the same main variable name, though with a different wave prefix. Since the last wave of BHPS was Wave 18, the wave prefix is "R". Thus if we wished to match Wave 2 work status (**b_jbstat**) on the file **b_indresp** to previous wave values, for the GPS sample we would match (using **pidp**) to **a_indresp** and use the variable **a_jbstat**, while for the BHPS sample we would match (using **pid**) to **rindresp** and use the variable **rjbstat**.

3.4. NOTES ON USING THE "EXTRA 5 MINUTES" QUESTIONS

An "Extra 5 minutes" of question time was set aside for questions of particular interest to ethnicity related research (e.g., ethnic identity and remittances). To provide sufficient sample sizes to allow analysis of these questions separately by ethnic (minority) groups, the questions were asked of the EMB sample (see Section 2.2.3), the GPC sample (see Section 2.2.2), and ethnic minority individuals living at Wave 1 in an LDA (i.e., members of ethnic minority groups who had a selection probability of 0 to be in the EMB sample). Those eligible for the "Extra 5 minutes" can be identified using ***w*_xtra5min_dv**. They also received the standard questionnaire that the rest of the sample received. The flag ***w*_xtra5minosm_dv** identifies OSMs who are eligible for these questions. In other words, ***w*_xtra5min_dv** is 1 and ***w*_xtra5minosm_dv** is 0 for TSMs who have joined the households of these OSMs who are eligible for the "Extra 5 minutes" questions.

To analyse this sample use the appropriate "Extra 5 minutes" (cross-sectional and longitudinal) weights (see Section 3.7 for further details on the construction of these and other weights). Note that there is no cross-sectional weight for the "Extra 5 minutes" questions as at Wave 2 these were only asked of sample members who had completed the main interview at Wave 1. Thus, the Wave 2 longitudinal weight should be used for Wave 2 cross-sectional analysis.

For further details on ethnicity analysis using UKHLS data see *Understanding Society*: Guide to Ethnicity Research (forthcoming).

3.5. DERIVED VARIABLES

Derived variables are variables that are computed from one or more variables. Some are computed by the Blaise CAPI program during the interview to control the routing within the questionnaire, whilst others are computed post-field for the purpose of analysts. The suite of derived variables included in the UKHLS includes flags for whether or not a certain characteristic is true for a study member (e.g., ***w*_jbft_dv** is a flag for whether or not a respondent has a full-time job), counts of the number of people in the household for whom a certain characteristic is true (e.g., ***w*_nemp_dv**

is the number of employed people in the household), and pointers to significant others in the household (e.g., **mnpid** records the cross-wave person identifier of the respondent's biological mother). As a rule of thumb, variables that are derived post-field end on the suffix "_dv", and pointers to others in the household end on "pno" or "pid"; they can, therefore, easily be identified on the data.

Note that a data file may offer alternative versions of a derived variable. This is particularly true for derived variables that point to others in the household: One set of variables (e.g. variable names starting on "hg") has been computed based on information collected in the household grid module of the questionnaire during the interview. The alternative version is computed post-field after the information collected in the household grid has undergone extensive data cleaning. See and compare, for example, ***w*_hgbiom** and ***w*_mnpno** for the person number of the respondent's biological mother in the household.

From Wave 2 onward, proactive dependent interviewing was used to increase efficiency of data collection and lessen respondent burden. Specifically, information reported at an earlier time is fed forward to the respondent to personalize the question. So rather than ask a question about current occupation with its complex probing by interviewers, the question might say, "the last time you were interviewed you said you were "specific occupation" are you still "specific occupation"? Feed-forward variables are used at both the household and individual levels. For example **b_ff_hhsize** feeds forward the household size from the previous wave (Wave 1). The variable **b_ff_plbornc** is the country of birth of the respondent fed forward from the previous wave. Note the use of the prefix "ff_". Some of the fed-forward variables were not used in the wording of a question but were used by the CAPI script to route respondents appropriately based on information from the previous wave.

Information collected using dependent interviewing is merged with the respective information collected using independent interviewing (e.g. when a respondent did not provide the information in the previous interview, or when they are new to the study) and stored in the data file under the variable name used for the latter (i.e., the variable stem name from Wave 1). See, for example, the socio-economic classification of the current job (***w*_jbsoc00**) and the standard industrial classification (***w*_jbsic07**).

We use look-up files between SOC 2000 and other classifications to derive additional occupational classifications. For further information and to obtain look-up files, see http://www.cf.ac.uk/socsi/CAMSIS/occunits/distribution.html#UK. We provide the following classifications International Standard Classification of Occupations (ISCO88), Registrar General Social Class (RGSC), National Statistics Socio-economic Classification (NS-SEC), Employment Status (ES), and Socio-economic Group (SEG). These are computed for the respondent's current job and last jobs only. SOC 2010 is provided but SOC 90 is not available for the current and last job.

Information about how the derived variable is produced is shown in the notes for derived variables in the detailed variable view of the online documentation. The view provides descriptive statistics and, in the Origin field, lists of the variables used in the computation of the derived variable. For variables that were computed during the interview, additional information is available in the questionnaires. Analysts should

consult the description of any (derived) variables that they plan to use in their analysis.

3.6. Sample Design Variables and Analysis

As the sample design involves stratification, clustering and weighting, these design features affect standard errors and should therefore be taken into account in analysis. Appropriate variables are provided to allow the analyst to do this. The weighting variables are described in Section 3.6. Here we describe the stratification and clustering variables.

3.6.1. Primary Sampling Units (PSUs)

***w*_psu** This is an indicator of the primary sampling unit (PSU) to which the sample member belongs. The prefix "*w*_" denotes waves in general.

The value of ***w*_psu** does not change between waves, but for new sample entrants it is only defined from the wave at which they enter the sample. Values on the variable ***w*_psu** are further described in Table 21.

Table 21: Description of the UKHLS Primary Sampling Unit variable (*w*_psu)

Value	Sample	Notes
1 – 575	former BHPS sample in England, Scotland and Wales	Identical to the BHPS variable ***w*psu**
701 – 1999	former BHPS Northern Ireland sample	Corresponds to initial (BHPS wave 11) sampled households, as these were selected in a one-stage design
2001 – 4640	UKHLS-GPS in England, Scotland and Wales	Corresponds to the postal sectors used as PSUs, see Lynn (2009)
46424 – 7035	UKHLS-GPS in Northern Ireland	Corresponds to Wave 1 sampled households, as these were selected in a one-stage design
7048 – 51789	UKHLS-EMB	Corresponds to Wave 1 sampled households, as these were selected in a one-stage design within the high minority density domain, see Berthoud, Fumagalli et al. (2009)

[Note: there was an error in **b_psu** and **c_psu** for Northern Ireland BHPS households in the Wave 2 and Wave 3 data releases. This has been corrected from the Wave 4 release.]

3.6.2. STRATA

***w*_ strata** This indicates the sampling stratum from which the sample member was selected.

The value of ***w*_strata** does not change between waves, but for new sample entrants it is only defined from the wave at which they enter the sample. The range of values on ***w*_strata** is listed in Table 22.

Table 22: Description of the UKHLS stratification variable (*w*_strata)

Values	Sample	Notes
1 – 151	former-BHPS sample in England, Scotland and Wales	Identical to the BHPS variable ***w*strata**
701	former-BHPS Northern Ireland sample	Northern Ireland treated as a single stratum
2001 – 3320	UKHLS-GPS in England, Scotland and Wales	Corresponds to groups of two or more PSUs in selection order, as they were selected systematically from an implicitly ordered list, see Lynn (2009)
3321	UKHLS-GPS in Northern Ireland	Northern Ireland treated as a single stratum
3322 – 5117	UKHLS-EMB	Corresponds to the postal sectors in the high minority density domain as selections were made independently from each, see Berthoud, Fumagalli et al. (2009)

3.6.3. ANALYSIS EXAMPLE USING STATA

In Stata, to obtain estimates that correctly take into account the sample design the user need only specify the design variables using the svyset command, for example:

```
svyset a_psu [pweight=a_indpxus_xw], strata(a_strata)
```

Then, any compatible command simply needs to be prefixed with “svy”, for example:

```
svy: logistic depvar  variable1  variable2  variable3
```

3.7. WEIGHTING ADJUSTMENTS FOR THE WAVE 4 RELEASE

A number of weights are provided for data users. They adjust for unequal selection probabilities, differential nonresponse, and potential sampling error. A weighted analysis will adjust for the higher sampling fraction in Northern Ireland and for different probabilities of selection in the EMB sample, as well as for response rate differences between subgroups of the sample. Separate sets of weights are provided for a) the combined GPS and EMB sample components, b) the former-BHPS sample, and c) the combined GPS, EMBS and BHPS components. The available sets of weights are not identical for these three analysis bases, reflecting differences in data collection. Considering the complexity of the study design, weights should be selected carefully, following the advice provided below.

The first part of this section covers the purpose of the weights and how to use the naming conventions for the weight variables to interpret and select the different weight variables from among a complex assortment. **This is the most important section for a user wanting to select the appropriate weight for a planned analysis.** This section is followed by the technical details of how weights were calculated.

If your aim is to generalise to the UK population, unweighted analyses should be avoided. For advanced users who want to model nonresponse in their own way, we provide design weights (see below) which adjust the sample for unequal selection probabilities. Note that adjusting for the first wave nonresponse is different from adjusting for attrition and requires variables which have values for both responding households and never responding households.

Note that a number of longitudinal weights are provided corresponding to the year of sample selection (BHPS since 1991; BHPS since 2001; GPS and EMB since 2010-2011; and combined BHPS, GPS and EMB since 2011-2012). Cross-sectional weights starting at Wave 3 will be based on combined BHPS, GPS and EMB samples.

3.7.1. SELECTING THE CORRECT WEIGHT FOR YOUR ANALYSIS

Given the complexity and multi-purpose nature of the UKHLS design we provide multiple weights to meet the different needs of users. The weight for your analysis reflects the survey instrument, which is the source of the data being used in the analysis, the analysis level (household or individual), and the combination of waves involved.

Each weight has been scaled to have a mean of one amongst cases eligible to receive the weight.

The naming conventions for weights are intended to help users to pick the correct weight. The name of each weight reflects the wave for which the weight is calculated, level of analysis, data source and its nature (design weight, cross-sectional analysis weight or longitudinal analysis weight). The rules are described in the “Naming Conventions for Weighting Variables” section below.

If your analysis uses only data from Wave 4, select the “xw” (cross-sectional) version of the weight. This weight is defined for all sample members who responded to the relevant survey instrument at Wave 4. If your analysis uses data from multiple waves select an appropriate “lw” (longitudinal) version of the weight.

For individual level analysis you may want to combine information from different questionnaire sources. In this situation please select the weight suitable for the lowest level according to the hierarchy below:

Table 23: Selecting the correct weight: Hierarchy of analysis levels

Level of Analysis	Questions available for
4	household level (all enumerated individuals)
3	Adult proxy and main interview
2	Adult main interview only (no proxy)
1	Adult or youth self-completion interview

For example, if in one cross-sectional model for Wave *n* you use questions from the proxy and full interview as well as from the self-completion, then the correct weight will be *n*_indscus_xw – the weight for the self-completion questionnaire, as its level (1) is lower than the level for proxy and full interview (3).

The following tables list the weight variables. The list has been broken into separate tables so the user can go quickly to the data source for the planned analysis and then select the particular relevant weight, for example cross-sectional vs. longitudinal. Each table focuses on a major data source and has the weight variables used for cross-sectional and longitudinal analyses. Please also note that weights are defined for particular sample components. Wave prefixes refer to a specific wave (a_ or b_) or to waves in general *n_.*

Here is an example: a longitudinal weight for BHPS + EMB + GPS represents waves 2 and 3, but the first time one can find it is in Wave 3 release. Thus, waves representing is 2-3, but the wave it starts from is 3. We would indicate that the above weight represents waves 2+ (all waves starting from 2).

Where is this useful? If a user wants to analyse longitudinal data starting with Wave 1, then they cannot use the combined longitudinal weight (for BHPS + GPS + EMB), but will have to use the weight for GPS +EMB (by looking at 'wave representing' they should be able to find quickly which weight is best for their analysis).

Table 24: Weight variables for analyses using household grid or household interview

Analysis level	Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
Household				
	1		Household grid and/or household interview	a_hhdenus_xw
	n	2	Household grid and/or household interview (BHPS, GPS and EMB)	*n*_hhdenub_xw
Individual				
	1		Household grid and/or household interview	a_psnenus_xw
	n	2	Household grid and/or household interview (BHPS, GPS and EMB)	*n*_psnenub_xw
	1+	2	Household grid and/or household interview (GPS and EMB)	*n*_psnenus_lw
	1991+	2	Household grid and/or household interview (BHPS, GB 1991)	*n*_psnen91_lw
	2001+	2	Household grid and/or household interview (BHPS, UK 2001)	*n*_psnen01_lw
	2+	3	Household grid and/or household interview (BHPS, GPS and EMB since 2010-2011)	*n*_psnenub_lw

Table 25: Weights for analysis using adult main and proxy interviews

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult main and proxy interview	a_indpxus_xw
2		Adult main and proxy interview	b_indpxus_xw
2		Adult main and proxy interview (BHPS)	b_indpxbh_xw
n	3	Adult main and proxy interview (BHPS, GPS and EMB since 2010-2011)	*n*_indpxub_xw
1+	2	Adult main and proxy interview	*n*_indpxus_lw
2+	3	Adult main and proxy interview (BHPS, GPS and EMB since 2010-2011)	*n*_indpxub_lw

Note: The weights listed in Table 25 through

Table 30 all apply to analysis at the individual level.

Table 26: Weights for analysis using adult main interviews

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult main interview	a_indinus_xw
2		Adult main interview	b_indinus_xw
2		Adult main interview (BHPS)	b_indinbh_xw
n	3	Adult main interview (BHPS, GPS and EMB)	*n*_indinub_xw
1+	2	Adult main interview	*n*_indinus_lw
1991+	2	Adult main interview (BHPS, Great Britain)	*n*_indin91_lw
2001+	2	Adult main interview (BHPS, UK)	*n*_indin01_lw
2+	3	Adult main interview (BHPS, GPS and EMB since 2010-2011)	*n*_indinub_lw

Table 27: Weights for analysis using adult "Extra 5 minutes" interview

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult extra 5 minutes interview	a_ind5mus_xw
n	3	Adult extra 5 minute interview	*n*_ind5mus_xw
1+	2	Adult extra 5 minutes interview	*n*_ind5mus_lw

Table 28: Weights for analysis using adult self-completion

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult self-completion	a_indscus_xw

2		Adult self-completion	b_indscus_xw
2		Adult self-completion (BHPS)	b_indscbh_xw
n	3	Adult self-completion (BHPS, GPS and EMB)	*n*_indscub_xw
1+	2	Adult self-completion	*n*_indscus_lw

Table 29: Weights for analysis using youth self-completion

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Youth self-completion	a_ythscus_xw
2		Youth self-completion	b_ythscus_xw
2		Youth self-completion (BHPS)	b_ythscbh_xw
n	3	Youth self-completion (BHPS, GPS and EMB)	*n*_ythscub_xw

Table 30: Weights for analysis using nurse health assessment data

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
2+	3	Nurse visit with Wave 2 to Wave *n* full interviews (GPS, BHPS)	*n*_indnsub_lw
1991+	3	Nurse visit with 1991 to Wave *n* full interviews (BHPS)	*n*_indns91_lw
Note: Other weights for use with the health assessment data that were collected at waves 2 and 3 are available with the separate health assessment data release, see Section 4.3.2, and are fully documented in the associated user guide, see McFall, Petersen et al. (2014).These two weights are documented here as they will be released with the main data release each year.			

Table 31: Design and inclusion weights*

Analysis level	Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
Household				
			(Wave 1 household design weight)	a_hhdenus_xd
			(Wave 1 household design weight for GPS only)	a_hhdengp_xd
individual				
			(Design weight)	a_psnenus_xd
			(Design weight GPS only)	a_psnengp_xd
			(Extra 5 minutes design weight)	a_ind5mus_xd
			BHPS inclusion weight for OSMs issued into UKHLS	b_psnenbh_li
			BHPS-2010 longitudinal enumerated person weight	b_psnenbh_lw
			BHPS, GPS and EMB combination inclusion weight	b_psnenub_li

*Typically for advanced users (see technical details)

3.7.1.1. Not Using Weights

Note, that an unweighted analysis does not correctly reflect the population structure unless the assumptions below are true. It is suggested that researchers publishing or presenting unweighted estimates make these assumptions explicit.

If no weighting is used, an analysis of the UKHLS assumes:

- that all estimates of interest are the same in Northern Ireland as in the rest of the UK;
- that people who live at an address with more than three dwellings or more than three households are the same as those who don't;
- that people who responded at Wave 1 are the same with respect to your estimates as those who did not; that people who continued to respond at later waves are the same as those who did not; and that people who responded to each particular instrument used in the analysis (individual interview, self-completion questionnaire etc) are the same as those who did not, see Lynn, Burton et al. (2012).

An unweighted analysis of the former-BHPS sample assumes:

- that all estimates of interest are the same in each of England, Scotland, Wales and Northern Ireland;
- that people who live at an address with more than three dwellings or more than three households are the same as those who don't;
- that people who responded at Wave 2 of UKHLS in 2010 are the same with respect to your estimates as those who may have become non-respondents at any time since Wave 1 of BHPS in 1991;
- that people who keep responding in later waves of UKHLS are the same as those who stopped responding at any point of time between 1991 and the last year in your analysis.

We therefore strongly suggest conducting weighted analyses of the UKHLS data.

3.7.2. Naming Conventions for Weighting Variables

The naming conventions will help users to select the weight they need or to interpret the purpose of some weight variables. The structure is as follows:

w_xxxyyzz_aa, where
w=wave
xxx=target population
yy=instrument
zz=sample
aa=weight type

options for xxx:
hhd: household
psn: persons 0+
ind: persons 16+
yth: persons 10-15

options for yy:
en: enumeration (grid)
in: interview
px: interview or proxy
5m: “extra 5 minutes” items
sc: self-completion
ns: nurse visit
bd: blood

options for zz:
us: the GPS and EMB samples
bh: BHPS sample
ub: BHPS, GPS and EMB samples
gp: GPS sample
91: BHPS original sample starting in 1991 (England, Scotland and Wales)
01: BHPS sample starting in 2001 (original sample + Scotland and Wales boost + NI)
ip: Innovation Panel

Options for weight type aa:
lw: longitudinal analysis weight
xw: cross-sectional analysis weight
ld: longitudinal design weight
xd: cross-sectional design weight
li: longitudinal inclusion weight

3.7.2.1. Examples

a_indinus_xw is the cross-sectional analysis weight for individual interview data from Wave 1, representing the population of persons aged 16 or older.

b_indscus_lw is the longitudinal analysis weight for individual self-completion interviews from Wave 1 and Wave 2 representing the adult population who continuously lived in UK at the times of Wave 1 and 2.

3.7.3. TECHNICAL DETAILS

In this section we describe in turn how the weights were derived for:

- GPS and EMB Wave 1 weight;
- GPS and EMB longitudinal weights;
- GPS and EMB cross-sectional weights after Wave 1;
- BHPS longitudinal weights;
- BHPS cross-sectional weights;
- Combined sample (BHPS, GPS and EMB) longitudinal weights;
- Combined sample (BHPS, GPS and EMB) cross-sectional weights.

3.7.3.1. Wave 1 (GPS and EMB) Weights

The Wave 1 household level weights consist of two components: a design weight and nonresponse adjustment for household level nonresponse. Wave 1 individual

level weights consist of four components: the design weight, nonresponse adjustment for household level nonresponse, individual level within-household nonresponse, and post-stratification to population characteristics. Each of the components is explained below.

Design Weight

The design weight corrects for unequal probability of selection at a number of levels.

The household level design weight corrects for:

- Unequal selection probability due to the boost in Northern Ireland. The GPS selection probabilities in Northern Ireland are approximately twice those in other parts of the UK;
- Unequal selection probability related to selection into the EMB. Selection probabilities in the EMB part of the sample vary considerably between areas, depending on the estimated ethnic mix of the area and ethnic composition of the household. Additionally, households in high density areas with at least one ethnic minority member were weighted to account for combined probability of being selected as part of the GPS or as part of the EMB samples;
- The selection probability of households in a dwelling with more than three households or at an address with more than three dwellings is adjusted for the fact that only three such households were selected from the same address.

Individual level design weights correct for all the above with one specific difference: non-EM persons who live with EM persons in the same household have a chance to be selected only via the GPS part of the sample, and not via the EMB. This means that non-EM persons in the EMB (who are TSMs) are given a design weight of 0 while non-EM persons in the GPS are given the household design weight. The weights for EM persons are adjusted for their dual probability of being part of GPS and EMB.

Individual level design weights for those eligible to answer the “Extra 5 minutes” is similar to the above design weight but differs in the following ways. It adjusts for the fact that the GPS-Comparison Sample is only 1/45th of the GPS original sample; that all EM members in low-density areas were administered the “Extra 5 minutes”; and that EM members in high-density areas had a chance to be selected into either the GPS Comparison sample or the EMB. Similar to the above weight, non-EM persons were assumed to have a chance to be part of the GPS Comparison Sample only and not part of the EMB.

Additionally, we provide GPS design weights (**a_hhdengp_dw** and **a_indengp_dw**). These weights are valid only for sample members selected through the GPS, and adjust for oversampling in Northern Ireland and for subsampling within households from multiple dwellings per address or multiple households per dwelling.

Household-level Nonresponse Adjustment

Household level nonresponse adjustment is more complex than in other surveys given the large number of households which were selected as part of the EMB with unknown eligibility. Households who were selected as part of the EMB sample were screened on whether they contain at least one member of a relevant ethnic minority group (Berthoud, Fumagalli et al. 2009). Given the low proportion of eligible households in the EMB sample it is unrealistic to assume that all non-responding

households would be eligible, that is, contain at least one EM member. To take this into account we modelled eligibility and used this information in household nonresponse adjustments such that households which were more likely to be eligible had a higher influence on the nonresponse correction. Note, that the predicted eligibility multiplied by the design weight is released for all the EMB sample households of unknown eligibility as part of **a_hhdenus_xd**. This will enable an advanced user to model Wave 1 household nonresponse taking into account the chance to be eligible among households of unknown eligibility.

To model eligibility we used predictors from the sampling frame and administrative neighbourhood data linked at a geographical level (for detailed description see below). After excluding ineligible addresses (like businesses or demolished and non-existent addresses), the eligibility was modelled using only EMB households with known eligibility status (either screened out or screened in). This prediction was then extrapolated onto EMB households of unknown eligibility (not contacted). Given the limited number of selected addresses in Wales and Scotland and differences between countries in the available auxiliary variables (see below), we predicted eligibility using two models. The first included common predictors for England and Wales and eligibility was predicted for these two countries. The second was based on England, Wales and Scotland, using a more limited number of predictors. Eligibility for Scotland was predicted only from this model.

Following this, the probability of responding was estimated using backward stepwise logistic regression weighted by eligibility status (where the ineligible were excluded, those known to be eligible had an eligibility of one, and those with unknown eligibility had a weight proportional to the predicted probability of being eligible obtained from the above model). The predictors used in this model were the same as for the eligibility model and are described in detail below. Given that administrative neighbourhood data differs between England, Wales, Scotland and Northern Ireland, a separate model was implemented for each country. GPS and EMB response propensity was modelled together (which allowed us to model nonresponse within each country separately), but the indicator of EMB was retained in the model even if it was not statistically significant.

Predictors used for eligibility model and household level nonresponse correction come from the following sources:

- Sampling frame information, including such variables as sample month and geographical region;
- Predicted ethnic density of the postcode sector for five main ethnic groups in England, Scotland and Wales, as described in Berthoud, Fumagalli et al. (2009);
- A wide range of indicators from Census 2001 and the most updated version of neighbourhood statistics as of summer 2011, linked separately for England, Wales, Scotland and Northern Ireland (see below).

The household nonresponse correction weight was calculated as the inverse of probability from the above model. This weight was multiplied by the household design weight to create the Wave 1 household level weight. The design effect was estimated using this weight. No truncation was necessary. The obtained weight was scaled to a mean of 1 and was named **a_hhdenus_xw**.

Neighbourhood Statistics

For England and Wales the information was linked at Middle Layer Super Output Area (MSOA) or Lower Layer Super Output Area (LSOA) levels and was obtained from http://neighbourhood.statistics.gov.uk. The examples of linked information obtained from Census 2001 include the proportions in the MSOA of employed, retired, outright property owners, travellers to work using different types of transport, single household members, households with one car, people with different types of qualification and professional occupation, among others. Other linked information includes 2010 information on multiple deprivation indexes, on crime instances, 2009 information on inflow and net change of neighbourhood population, the proportion of different allowance claimants, and 2008 information on hospital admissions and energy consumption.

For Scotland the information was linked at the data zone level from http://www.scrol.gov.uk/scrol/common/home.jsp and from http://www.scotland.gov.uk/Topics/Statistics/SIMD. From the Census 2001, information was obtained on population density, mean age, average household size, and number of rooms per household in the data zone, as well as the proportions in the data zone born in Scotland and outside the EU, of different religious denominations, employed, unemployed and retired, disabled, those with different levels of qualification and types of occupation, and different types of accommodation, among others.

For Northern Ireland the information was linked at the Super Output Area (SOA) level and was obtained from http://www.ninis.nisra.gov.uk/. Examples of predictors obtained from Census 2001 at the SOA level include the average hours worked by residents, the average age of residents, percentages of residents with different level of qualifications, with different employment statuses, and with different types of marital status, among others. The predictors also include 2007-2009 information on multiple deprivation indexes.

Note, that using *Understanding Society* analysis weights (all but design weights), adjusts for household nonresponse bias in any estimate, to the extent it is related to the above mentioned variables.

Enumerated Individual Weight

The weight for analysis of enumerated individuals (**a_psnenus_xw**) is not equivalent to the household weight for all household members, as often happens in other household studies. This is because we have TSMs in Wave 1, who are not ethnic members, selected into EMB part of the sample. Thus, the individual level design weight is not equal to the household level design weight for individuals in households containing a mix of EM and non-EM persons. The weight for the analysis of enumerated individuals is calculated as the product of the individual level design weight **a_psnenus_xd** and the household level nonresponse correction (described above). The design effect was tested showing that no truncation was necessary. Weighted sample distributions were then compared to ONS mid-year estimates (with a correction for institutionalized population) and post-stratification was implemented for the fully crossed matrix of gender by geographical region by 5-10 year age groups. Thus the individual level enumerated weight consists of:

The individual level design weight * the household nonresponse correction * the post-stratification adjustment. The obtained weight is scaled to have a mean of one.

Individual-level Nonresponse Adjustment

Five different individual level weights were prepared for users reflecting nonresponse occurring at different levels and different questionnaire instruments. Each individual level weight consists of this product:

The individual level design weight * the household nonresponse correction * the individual level nonresponse correction conditional on household response * the post-stratification adjustment.

The individual nonresponse correction (conditional on household nonresponse) is modelled at three levels:

- For adult respondents (age 16 or older) who either completed the main interview or for whom a proxy interview was completed (for **a_indpxus_xw**);
- For adult respondents (age 16 or older) who completed the main interview only (for **a_indinus_xw** and **a_ind5mus_xw**);
- For respondents aged 10 or older who completed and returned the self-completion questionnaire (for **a_indscus_xw** and **a_ythscus_xw**).

Note, that the same model was used for respondents regardless of whether they were selected into GPS or EMB; that the response propensity is assumed to not depend on whether respondents received the “Extra 5 minutes” or not; and that conditional on age (present in the model), the response to self-completion is assumed to have the same predictors for adults and youth (this assumption allowed modelling the response in each country separately, which wouldn’t otherwise be possible for youth sample).

The individual level response, conditional on household response, was modelled using backward stepwise logistic regression separately for England, Wales, Scotland and Northern Ireland. The four models were implemented for each of the three levels described above. The predictors used in the models include all the predictors used for the household level nonresponse models and individual and household-level variables obtained from the household questionnaire, such as age and gender, marital and employment status, household size and presence of children in the household, as well as household expenditure on food and food outside, consideration of use of environmental energy, among others.

The individual-level nonresponse adjustment was obtained as the inverse of the predicted probability and was then multiplied by the relevant (either individual or “Extra-five minutes”) design weight and by the household nonresponse correction. No truncation was deemed necessary as there were no extreme values substantially impacting the design effects. The post-stratification was implemented as described above in the individual level enumeration weight section, except that a greatly reduced matrix was used in the case of the “Extra-five-minutes” weight, due to the much smaller sample size for which this weight applies. After multiplying by the post-stratification adjustment, each of the obtained weights was then scaled to a mean of one.

3.7.3.2. GPS and EMB Longitudinal Weights

Each of the five types of longitudinal weights (enumerated persons, proxy or main interview, main interview, self-completion, and "Extra-five minutes" interview) is based on the corresponding previous longitudinal weight (except in Wave 2 where it is based on Wave 1 cross-sectional weight). An additional adjustment for nonresponse since the last wave is applied. Each adjustment is based on a model of Wave *n* response conditional on Wave (*n*-1) non-zero longitudinal weight for the instrument in question. For the enumerated person model, covariates are taken from the Wave (*n*-1) household grid and household questionnaire. In the model for proxy and main interviews, covariates were taken from the Wave (*n*-1) proxy interview (or the equivalent items from the main interview), household grid and household questionnaire. In both the model for main interviews and the model for adult self-completion questionnaires, covariates are taken from the Wave (*n*-1) main interview, household grid and household questionnaire. The adjustment weight is calculated as the reciprocal of the model-predicted response propensity. The Wave (*n*-1) weight is then multiplied by the Wave *n* adjustment to create the Wave *n* longitudinal weight.

Newborns born to an OSM mother since the Wave (*n*-1) interview receive the longitudinal enumerated person weight of their mother (reflecting the idea that the probability of observing the newborn is equal to the probability of observing the mother). The principle behind the longitudinal weights is that they are defined for each person who is observed at all of the relevant waves for which they were eligible. For this reason, newborns observed at Wave *n* receive a Wave *n* longitudinal weight as they were enumerated at Wave *n*, the only wave for which they were eligible.

3.7.3.3. Cross-sectional Weights (GPS and EMB) for Wave 2

The cross-sectional enumerated individual weights are based on the longitudinal enumerated individual weights, which are allocated through a weight-share method to temporary sample members (TSMs) and permanent sample members (PSMs) who entered the sample at Wave 2. Note, that only new TSMs and PSMs entering the study after Wave 1 receive a shared weight. TSMs who were present in Wave 1 (in the EMB sample) are given a cross-sectional weight of 0. This is done as the GPS part of the sample does not have an equivalent TSM group (OSM non-ethnic minority members living with TSM ethnic minority members). Giving a cross-sectional weight of 0 to Wave 1 TSMs maintains the balance of the whole sample.

These cross-sectional enumerated individual weights then serve as the base for the other cross-sectional individual-level weights, each of which (main, main or proxy, self-completion, youth) involves an additional adjustment for nonresponse to the relevant instrument conditional on enumeration. The nonresponse models are therefore based on all eligible persons enumerated at Wave 2 (including TSMs and those OSMs who did not respond to the respective instrument at Wave 1), with covariates taken from responses to the UKHLS Wave 2 household grid and household questionnaire.

The cross-sectional weights for households (**b_hhdenus_xw**) are set equal to the minimum nonzero longitudinal enumerated person weight (**b_psnenus_lw**) amongst adults in the household, reflecting the idea that the probability of observing the household is equal to (or greater than) the probability of observing the person in the household who has the greatest probability of being observed.

3.7.3.4. BHPS Longitudinal Weights

Four weights will be continued from BHPS with changed variable names. The corresponding weight variables are:

- ***w*ewght** now called ***w*_psnen91_lw**,
- ***w*lewtuk1** now called ***w*_psnen01_lw**,
- ***w*lrght** now called ***w*_indin91_lw**, and
- ***w*lrwtuk1** now called ***w*_indin01_lw**, where ***w*** represents the most recent UKHLS wave.

These weights are based on Wave 18 BHPS longitudinal weights, which account for the first wave household nonresponse, the first wave within household individual nonresponse (to enumeration or to an individual main questionnaire, respectively) and for individual nonresponse between the first wave and Wave 18 of BHPS. The base weights which reflect continuous enumeration (**rlewght**, a BHPS variable name) and continuous response to the main questionnaire (**rlrght**, a BHPS variable name) since 1991 are used for creating weights for longitudinal analysis starting 1991. Note that such an analysis excludes Northern Ireland as it was added to BHPS in 2001 and will also exclude the Scotland and Wales boost samples that were added in 1999. Similarly, the base weights which reflect continuous enumeration (**rlewtuk1**, a BHPS variable name) and continuous response to main questionnaire (**rlrwtuk1**, a BHPS variable name) since 2001 are used for creating weights for longitudinal analysis starting in 2001. Analysis using these weights will include all the BHPS samples. For more information on the BHPS weight calculation please refer to BHPS documentation (Taylor 2010).

For each of the Wave 18 weights an additional adjustment is applied to correct for attrition between Wave 18 of the BHPS and Wave 2 of *Understanding Society*, when the BHPS joined *Understanding Society.* The adjustment is the reverse of the estimated probabilities of participation (enumeration or response to main questionnaire) based on logistic regressions predicting participation at Wave 2 of UKHLS conditional on participation at Wave 18 of BHPS. The covariates used in the model predicting enumeration are from the BHPS Wave 18 household grid and household questionnaire. The same covariates plus covariates from the Wave 18 main questionnaire are used for predicting response to the UKHLS Wave 2 main questionnaire. Enumeration weights for newborn babies (biological, step or natural) born to an OSM mother since the time of the BHPS Wave 18 interview are equal to their mother's enumeration weight.

For "rising 16 year-olds" (OSMs who turned 16 between the time of the BHPS Wave 18 interview and the UKHLS Wave 2 interview and who could therefore be aged 16, 17, or even 18 at the time of UKHLS Wave 2), main response weights consist of the relevant longitudinal enumerated person weight, with an adjustment for the probability of main response at Wave 2 conditional on enumeration at Wave 2. The adjustment is the inverse of the response propensity predicted by a separate logistic regression model (based just upon all adults and inferred to rising 16 year-olds) using covariates from the Wave 2 household questionnaire and household grid. The base weight for rising 16-year olds correction is continuous enumeration since 1991 (**b_psnen91_lw**) for the BHPS 1991 main response weight (**b_indin91_lw**), and is the BHPS-2010 longitudinal enumerated person weight (**b_psnenbh_lw** – see next section below) for the BHPS 2001 main response weight (**b_indin01_lw**). The main

response weight for each rising 16 year-old is then scaled by a constant factor so that the ratio of rising 16 year olds to older adults among main questionnaire respondents equals the equivalent proportion among all enumerated respondents. The weights (**b_psnen91_lw**, **b_psnen01_lw**, **b_indin91_lw** and **b_indin01_lw**) are calculated by multiplying the respective BHPS Wave 18 weight and the adjustment, and are scaled to one.

Starting at Wave 3, longitudinal weights for BHPS are created based on the previous wave longitudinal weights. For longitudinal enumeration weights (***n*_psnen91_lw** and ***n*_psnen01_lw**), enumeration in Wave ***n*** is predicted among adults having positive longitudinal weight in previous wave. Enumeration is modelled using logistic regression with covariates from the household questionnaire and household grid from wave *n*-1. Newborns are given the enumeration weight of their mother. The enumeration weights are then scaled to a mean of one.

For longitudinal main interview weights (***n*_indin91_lw** and ***n*_indin01_lw**) response to the main interview is predicted using logistic regression with predictors obtained from the wave *n*-1 household questionnaire, household grid, and main questionnaire. The probability is then inverted. The response of rising 16 year olds is predicted in a separate logistic regression in which response to the main questionnaire for all adults 16 and over is estimated using predictors from the wave *n* household questionnaire and household grid and conditional on enumeration in current wave. The response probabilities are then inverted and multiplied by the base weights (***n-1*_indin91_lw** or ***n-1*_indin01_lw** for adults; and ***n*_psnen91_lw** or ***n*_psnen01_lw** for rising 16). The weighted ratio of rising 16 to others is scaled to reflect the ratio of these age groups as estimated using the longitudinal enumeration weight in wave *n.*

3.7.3.5. BHPS Cross-sectional Weights for Wave 2

The BHPS cross-sectional weights are created as follows: we first model the chance of each BHPS OSM being issued into the UKHLS (reflected in **b_psnenbh_li**), then the chance of being in a responding household (complete the household grid and the household questionnaire) at Wave 2 of the UKHLS conditional on being issued (reflected in **b_psnenbh_lw**). The weight **b_psnenbh_lw** is then extrapolated to TSMs and PSMs through a weight-share method to create **b_psnenbh_xw**. The detailed procedure for creating these weights as well as cross-sectional individual response weights is described below.

The inclusion weight (**b_psnenbh_li**) was calculated separately for a) Northern Ireland and b) England, Scotland and Wales. For each, it has two components. For Northern Ireland, the first component consists of the BHPS Wave 11 cross-sectional weight, as this is the wave at which Northern Ireland first entered the BHPS. This component encompasses a design weight, post-stratification and an adjustment for Wave 11 nonresponse. The second component is derived from a model of the propensity to be issued at UKHLS Wave 2 conditional on being enumerated in BHPS Wave 11. This therefore adjusts for all the stages of dropout between BHPS Wave 11 in 2001 and UKHLS Wave 2 in 2010. Model covariates were taken from the Wave 11 household grid and household questionnaire. This propensity was modelled as a single step from 2001 to 2010 because across-wave response patterns varied greatly between the sample members. There is no single BHPS wave since Wave 11 at which all the Northern Ireland sample members (of those issued to UKHLS)

responded and therefore no other survey instrument that can provide model covariates for all relevant sample members.

Similarly, for England, Scotland and Wales the first component consists of the BHPS Wave 9 longitudinal weight, as this is the wave at which the Scotland and Wales boost samples were added (so, all of the members of those samples who entered UKHLS were enumerated at that wave, as were the vast majority of members of the original BHPS Wave 1 sample who entered the UKHLS). This component therefore encompasses a design weight, Wave 1 post-stratification and adjustments for nonresponse at each of the Waves 1 to 9 of the BHPS. The second component is derived from a model of the propensity to be issued at UKHLS Wave 2, conditional on being enumerated in BHPS Wave 9. This adjusts for all the stages of dropout between BHPS Wave 9 in 1999 and UKHLS Wave 2 in 2010. Model covariates were taken from the Wave 9 household grid and household questionnaire.

BHPS OSM newborns since Wave 9 (England, Scotland or Wales) or Wave 11 (Northern Ireland) whose parents are both OSMs were then assigned a base weight equal to the smaller BHPS inclusion weight of their (OSM) parents in the child's 2010 (issued to UKHLS) household. This reflects the idea that the probability of the child entering the UKHLS sample equals the probability of at least one of his or her parents entering the sample, which in turn is equal to (or greater than) the probability of the parent who has the greatest probability of entering the sample. BHPS OSM newborns born to one OSM parent and one TSM parent were assigned a base weight equal to half of the OSM parent's weight in the child's 2010 (issued to UKHLS) household. The division by two reflects the idea that these newborns had double the chance of becoming BHPS OSMs, relative to people born to both OSM parents, as they would have been included had either their mother's or father's 1991 household been sampled. For newborns observed with a single parent in a household in the first wave after their birth, the weight given was the parent's weight. This reflects a close to zero likelihood for the baby to be sampled via the other parent.

The adjustment for household nonresponse at UKHLS Wave 2 was derived from a model of enumeration at Wave 2 conditional on entering the UKHLS sample (i.e. being issued to the field for UKHLS Wave 2), in which covariates came from the Wave 9 household instruments for England, Scotland and Wales and the Wave 11 household instruments for Northern Ireland. The weight which reflects the chance of a BHPS OSM of being selected into the BHPS, to be issued into UKHLS and to be enumerated at Wave 2 of UKHLS is the BHPS-2010 longitudinal enumerated person weight (**b_psnenbh_lw**).

Finally, the BHPS cross-sectional enumeration weight (**b_psnenbh_xw**) was created through a weight-share method by sharing the BHPS-2010 longitudinal enumerated person weight to TSMs and PSMs.

The BHPS cross-sectional weights for main, proxy or telephone interview respondents (**b_indpxbh_xw**), main interview respondents (**b_indinbh_xw**) and self-completion respondents (adults (**b_indscbh_xw**) and youth (**b_ythscbh_xw**) each consist of the cross-sectional individual enumerated weight with an additional adjustment for nonresponse to the relevant instrument conditional on household response. These adjustments were based on logistic regression models with both

individual-level and household-level covariates taken from responses to the UKHLS Wave 2 household grid and household questionnaire.

The BHPS cross-sectional household weight (**b_hhdenbh_xw**) is set equal to the minimum cross-sectional person enumerated weight (**b_psnenbh_xw**) amongst adults in the household.

3.7.3.6. Combined Sample (BHPS, GPS and EMB) Weights

Starting at Wave 3 we provide both cross-sectional and longitudinal weights for combined analysis of respondents in all samples, including BHPS, GPS and EMB. The weights are based on the inclusion enumeration weight, which accounts for combined probabilities of being selected in any of the continuing samples of BHPS, GPS and EMB at the time each was selected, and continuously being enumerated up to and including Wave 2 of UKHLS.

We first explain the calculation of the inclusion enumeration weight, the development of cross-sectional weights for Wave 2 based on it, then the calculation of longitudinal weights, and finally the cross-sectional weights that are created from Wave 3 onwards.

3.7.3.7. BHPS, GPS and EMB Inclusion Enumeration Weight

To combine samples from the BHPS, GPS and EMB sample components, we calculate the joint probability of each respondent being selected through each sample and to continue being enumerated up to and including Wave 2 of the UKHLS. Specifically, the following samples are combined:

Sample 1: BHPS 1991 (those living in England, Scotland and Wales)
Sample 2: BHPS 1999 (those living in Scotland and Wales)
Sample 3: BHPS 2001 (those living in Northern Ireland)
Sample 4: GPS 2009-2010 (those living in England, Scotland, Wales and Northern Ireland)
Sample 5: EMB 2009-2010 (those living in HDAs in England, Scotland and Wales)

Each respondent therefore had up to five chances to be selected into UKHLS, depending on where they lived at the time that each sample was selected. To reflect this, for each person (selected into any of the five samples) we calculate their probability of being sampled in each of the samples above (p_{jk}, *j* = 1, … , 5) and add these together to derive the overall inclusion probability ($p_{\bullet k}$):

$$p_{\bullet k} = \sum_{j=1}^{5} p_{jk}$$

The probability of respondent *k* entering UKHLS through sample *j*, p_{jk}, is calculated as:

$p_{jk} = s_{jk} \times h_{jk} \times i_{jk} \times v_{jk}$, where

- s_{jk} is the selection probability of respondent *k* in sample *j*;

- h_{jk} is the average household response rate at wave 1 for sample *j* in the country of residence of respondent *k* at the time of sample selection;
- i_{jk} is the average individual continuous enumeration probability (rate) between Wave 1 for sample *j* and Wave 2 of UKHLS, specific to respondent *k*'s country of residence at the time of selection of sample *j*. This is in effect to the inverse of the average individual attrition rate.
- v_{jk} is the individual-specific variation from the mean continuous enumeration probability. This is explained in detail below.

For the sample in which the individual was selected we know the selection probability and modelled response probabilities (these can be obtained from the available weights, and the method is described below). These probabilities will be called real. For the samples that the individual could have been selected through but wasn't, we can infer each of the components. Such probabilities are called inferred. The inference method is described below.

Selection probability (s_{jk})

For *j* = 1,2,3, s_{jk} for respondents known or inferred to have been resident in relevant country at the time of selection is equal to the number of eligible households divided by the number of residential households in the population at that time. Such probabilities are calculated separately for each country (England, Wales, Scotland and Northern Ireland). The population information was obtained from Census 1991 for the 1991 sample (Office for National Statistics 1991), and from Census 2001 for samples selected in 1999 and 2001 (Office for National Statistics 2003). For respondents known or inferred not to have been resident in relevant country at the time of selection, $s_{jk} = 0$.

For respondents actually selected into the GPS or EMB sample, the real probabilities are simply the inverse of the design weight (**a_hhdenus_xd**) the calculation for which is described above. Two inferred probabilities are calculated for respondents actually selected into the BHPS samples: the probability of selection through GPS (s_{4k}) and the probability of selection through EMB (s_{5k}). s_{4k} is equal to the number of eligible residential household selected through GPS divided by the number of households in the population according to Census 2011.– These probabilities are country specific.

s_{5k} depends on the postcode sector of residence in 2009-10 (as described in Berthoud et al, 2009), and on the ethnic composition of the household and the selected person. For BHPS sample members we assigned the EMB postcode sector probability of the postcode of their address at wave 2 of UKHLS (2010). If the sector was in a LDA, then $s_{5k} = 0$.

Next, for BHPS sample members in high density (ethnic minority) areas we need information on the ethnic group composition of the household. We used information from variables **race** and **racel** as recorded in data file **xwavedata** at the time of BHPS Wave 18, and **b_racel** as recorded at Wave 2 of the UKHLS. For children under 16, for which this information was missing, we inferred it from their biological parents. The information collected in the BHPS does not perfectly match the ethnic group classification used for EMB selection, but after consulting experts on migration

research we achieved a reasonable fit to the EMB classification. While the selection into EMB depended on whether a person belongs to the group of interest or not (see section on EMB), the selection probability depends on the household ethnic group composition at the time of selection (considered 2010 for this purpose). If the household did not have any ethnic minority member of interest, we set $s_{5k} = 0$. If the household had some ethnic minority members of interest, the probability was calculated as the product of postcode sector selection probability and the largest screening probability among ethnic group members. All ethnic minority members were assigned this selection probability, and members of any other groups (either British or those not of selected in EMB) were assigned the value of zero.

Additionally, postcode sectors from which the Bangladeshi boost was selected were identified and those of Bangladeshi origin were assigned a probability reflecting the additional chance of being selected into Bangladeshi boost, see Berthoud, Fumagalli et al. (2009).

Average household response rate at Wave 1 (h_{jk})

We treat nonresponse correction as having three components, the first of which is the household response probability at the first wave when the sample is selected. Here real and inferred probabilities are calculated in the same way. We divide the number of responding households by the number of eligible residential households in the sample – these are country and time specific. For the EMB sample the household response rate was calculated using the design weight to account for unknown eligibility.

Average individual continuous enumeration probability (i_{jk})

The second nonresponse component is an average individual continuous enumeration probability. This probability is calculated as the ratio of the number of people who were enumerated continuously in all waves since selection up to and including Wave 2 of the UKHLS to the number of people who were enumerated in the wave of selection minus those known to have become ineligible (deceased or out of scope). In other words this is the response rate (where nonresponse is defined as missing at least one wave).

Individual-specific variation from the mean continuous enumeration probability (v_{jk})

The third nonresponse component reflects the variability of enumeration propensities among different individuals. We obtain it by the following procedure: first, we invert the longitudinal enumeration weight for Wave 2 of UKHLS (**b_psnenus_lw** for the GPS and EMB samples and **b_psnenbh_lw** for the BHPS sample); second, we divide this inversed weight by three probabilities described above: selection probability (s_{jk}), average household probability of respond in wave 1 (h_{jk}), and average individual enumeration probability (i_{jk}). The remainder is then scaled to a mean of 1.00 within each country and each time point of selection, reflecting the enumeration probability for each person, relative to the average person selected at the same time in the same country. This is the real unique component available for the time points when the person was actually selected.

To obtain tv_{jk} for the time points when a person could have been selected but wasn't, we used the following procedure. For each country and time point we ran a

multiple stepwise regression with v_{jk} as the dependent variable and covariates from the household questionnaire and household grid from Wave 2 of the UKHLS as explanatory variables. Note that identical predictors are available for all samples of interest. We can therefore use the regression model to infer v_{jk} for sample members selected through another sample.

Residency

In order to assign a correct probability for each person we need to know where they lived in 1991 (England, Scotland, Wales or other), 1999 (Scotland, Wales or other), 2001 (Northern Ireland or other) and 2009/2010 (postcode sector if England, Scotland and Wales, Northern Ireland, other). Because no interviewing was conducted among BHPS members in 2009, we treat their address at the time of Wave 2 of the UKHLS (2010) as if it were their address at the time of selection of GPS and EMB.

We do not have perfect information for each enumerated respondent on their residency in the four years of interest, but we use all the available information to us to infer residency as closely as possible.

Specifically, the following information is used: the country and year of the respondents' selection; the respondent's report on the most recent change of address; for those not born in Britain, the year of their arrival to Britain the country and year of school or university studies. For adults whose residence remains unknown after using the above information, it is inferred from other household members, if it is consistent.

Note, that for those who were born after the sample was selected, we make use of residency information from their mother. In this way we obtain residency information with 5 categories (England, Wales, Scotland, Northern Ireland or abroad) for 1991, 1999, 2001 and 2009/2010 time points for each single respondent.

Total probability

The important point is that an estimate of each of the following probabilities is available for each person, regardless of when and where the person was selected:

p_{1k} if resident in England in 1991;
p_{1k} if resident in Scotland in 1991;
p_{1k} if resident in Wales in 1991;
p_{2k} if resident in Scotland in 1999;
p_{2k} if resident in Wales in 1999;
p_{3k} if resident in Northern Ireland in 2001;
p_{4k} if resident in England in 2009/10;
p_{4k} if resident in Scotland in 2009/10;
p_{4k} if resident in Wales in 2009/10;
p_{4k} if resident in Northern Ireland in 2009;
p_{5k} if resident in England in 2009/10;
p_{5k} if resident in Scotland in 2009/10;
p_{5k} if resident in Wales in 2009/10.

For each time point (1991, 1999, 2001, and 2009/2010) the probability is non-zero only for the sample reflecting the respondent's country of residence at that time, and

is zero for all other countries for that time point. Thus for a person who has always lived in England the non-zero selection probabilities will be those for England in 1991 and for England for 2009/2010. For those who immigrated to England in 2008, all selection probabilities will be zero except for the probabilities for England in 2009/2010.

The total probability, $p_{\bullet k}$, is therefore the sum of all the above probabilities. It reflects multiple possible ways of selection through different samples and continuous enumeration of a respondent up to and including Wave 2 of the UKHLS. The inclusion weight (**b_psnenub_li**) is the inverse of the total probability. This weight is not scaled and will serve as a base weight for all the weights that combine BHPS, GPS and EMB samples. The weight is defined for all OSM respondents enumerated at Wave 2 of the UKHLS.

3.7.3.8. W2 Cross-sectional Weights for Total Sample (BHPS, GPS and EMB Components)

Two cross-sectional weights are created for Wave 2 of the UKHLS for the combined samples of BHPS, GPS and EMB. The first is the cross-sectional individual enumeration weight which is created through the weight share technique where the inclusion weight (**b_psnenub_li**) is weight-shared to TSMs and those OSMs that have missed at least one wave between selection and Wave 2 of the UKHLS. The weight share is done in a standard way: the inclusion weight, **b_psnenub_li**, is unaltered if it is non-zero for all household members but the cross-sectional weight is equal to the average inclusion weight for households where at least one member has a value of zero for the inclusion weight. The obtained cross-sectional enumeration individual weight is then scaled to a mean of one (**b_psnenub_xw**).

For each household the lowest **b_psnenub_li** from OSM adults (16 years of age or older) is selected. This reflects the highest probability of enumeration among OSM household adult members. The weight is then scaled to the mean of one (**b_hhdenub_xw**).

3.7.3.9. Longitudinal Weights for Total Sample (BHPS, GPS and EMB Components) for Waves 3 and 4

A longitudinal enumeration weight for the BHPS, GPS and EMB samples (**c_psnenub_lw**) was created based on the inclusion enumeration weight into Wave 2 (**b_psnenub_li**). Conditional on a nonzero value for **b_psnenub_li**, the enumeration is modelled using logistic regression and predictors from the wave 2 household questionnaire and household grid. The estimated probabilities of conditional enumeration are inversed, and multiplied by **b_psnenub_li**. Newborns are given their mother's enumeration weight. The weight is then scaled to a mean of one.

Longitudinal response weights for proxy and main interview, main interview, and self-completion interview were also created at Wave 3. All of these are based on longitudinal enumeration in Wave 3 (positive value of **c_psnenub_lw**). For each instrument, a logistic regression is run to predict response in both Waves 2 and 3 with predictors from the Wave 3 household questionnaire and household grid. The models are restricted to adults. The estimated probabilities were then inversed, multiplied by **c_psnenub_lw** and scaled to a mean of one.

From Wave 4 onwards, each of these longitudinal weights are based on the equivalent weight from the previous wave, adjusted by the reciprocal of the predicted value from a logistic regression model of wave-on-wave response and scaled to a mean of one. For example, a logistic regression model of enumeration at Wave 4 was based on sample members with a non-zero value of **c_psnenub_lw** and the reciprocal predicted values were multiplied by **c_psnenub_lw** and then scaled to produce **d_psnenub_lw**.

3.7.3.10. Cross-sectional Weights for Waves 3 and 4

Starting at Wave 3, cross-sectional weights were created for the combined sample that includes the BHPS, GPS and EMB components. The cross-sectional enumeration weight (***n*_indenub_xw**) was created based on the longitudinal enumeration weight (***n*_indenub_lw**) via the weight-share method. The weight was shared from OSMs with nonzero longitudinal enumeration weights to TSMs and PSMs (except those who were selected at Wave 1), and OSMs with a longitudinal weight of zero.

For the household cross-sectional weight (***n*_hhdenub_xw**) the lowest cross-sectional enumeration weight among adults (***n*_indenub_xw**) was selected. The weight was then scaled to a mean of one.

Cross-sectional weights for proxy and main (***n*_indpxub_xw**), main (***n*_indinub_xw**) and self-completion (***n*_indscub_xw**) were created based on the Wave *n* cross-sectional enumeration weight. Three logistic regressions were run to predict the response for the relevant instrument with predictors from the household grid and household questionnaire in Wave *n*, conditional on enumeration in Wave *n*. The models were restricted to adults. The estimated probabilities were inversed, multiplied by ***n*_indenub_xw** and were scaled to a mean of one.

The cross-sectional weight for youth (***n*_ythscub_xw**) was also created based on enumeration at the same Wave *n* (***n*_indenub_xw**). Among eligible youth members (between 10 and 15) a logistic regression was run with predictors from the household grid and household questionnaire from Wave *n* to predict response to the youth questionnaire. The inverse of the probability was multiplied by ***n*_indenub_xw** and was scaled to a mean of one.

3.8. IMPUTATION OF INCOME VARIABLES

Understanding Society collects detailed information each wave on personal income. All individuals aged 16 or more are asked to report:

- wages,
- self-employment earnings,
- second job earnings,
- interest and dividends,
- pensions (National Insurance/state retirement pension, pension from a previous employer, pension from a spouse's previous employer, private pension/annuity, widow's or war widow's pension, widowed mother's allowance or widowed pension),
- benefits (severe disablement allowance, disability living allowance, war disablement pension, attendance allowance, carer's allowance, incapacity

benefit, income support, job seeker's allowance, national insurance credits, child benefit, child tax credit, working tax credit, maternity allowance, housing benefit, council tax benefit, foster allowance/guardian allowance/rent rebate, rate rebate, employment and support allowance, respond to work credit, sickness and accident insurance, in-work credit for lone parents and pension credit) and

- other income sources (educational grant, trade union and friendly society payment, maintenance or alimony, payments from a family member not living together, amount for rent from boarders or lodgers, rent from any other property).

These personal income variables can be summed to obtain the total personal income. Total household income can be computed from the personal total incomes of all household members.

Some of the income components can be missing. More precisely there can be three types of missing cases:

- item nonresponse when individuals respond to the individual questionnaire but do not answer to some or all the questions on income components;
- individual nonresponse when individuals fail to respond to the individual questionnaire;
- household nonresponse when there is neither a household nor the individual questionnaire response.

For example, at Wave 1 we have 59,466 individuals for whom at least the household questionnaire is available, and, among these individuals, 80.3% provided a personal interview, 5.5% have a proxy interview, whereas 14.2% had neither a proxy nor a personal interview. The item nonresponse rate for individuals who provided an individual questionnaire varies across income variables. It goes from a maximum of about 50% for self-employment earnings to zero for some of the benefit variables, and it is generally below 20% for the remaining income variables.

3.8.1. WHAT DO WE IMPUTE?

In *Understanding Society* we do not impute income variables for non-responding households. Responding households are households for which the household questionnaire and information on the household composition/structure (household grid module) are available. We suggest that the user take account of household nonresponse via weighted estimates, described in Section 2 Weighting adjustments).

For individuals who respond to the individual questionnaire but do not provide answers to all income questions (item nonresponse), we impute the following personal income variables: wages, self-employment earnings, second job earnings, interests and dividends, pensions, benefits and other income sources.

For individuals for whom a proxy questionnaire is available, we impute total earnings and total income whenever missing. The proxy questionnaire is a short version of the individual questionnaire with questions on total earnings and total income as well as other variables.

Finally, for individuals in responding households for whom neither the personal nor the proxy questionnaire is available, we impute only the total personal income. This

is not directly included in the data set, but is used in the imputation of total household income.

Based on these imputations we can compute total personal and household income for all individuals belonging to responding households.

For each income variable for which amounts are imputed there is a separate imputation flag variable (with a suffix "_if" instead of "_dv" indicating whether the variable is imputed. In most cases this takes the value 1 if imputed and 0 if not, but in the case of the following variables it shows the proportion of total income imputed: ***w*_fimngrs_if**, ***w*_fibenothr_if**, and ***w*_fihhmngrs_if**.

In the income data file there may be multiple receipts of income from the same source. For example, a respondent may have multiple pensions from a previous employer. These are summed and imputed as such, and the imputed values are in the variable ***w*_frmnthimp_dv**. As a consequence, the variable ***w*_frmnthimp_dv** for the first income receipt from a given source is equal to the total value of all receipts from that source, while it is set to inapplicable for the second and subsequent receipt.

3.8.2. IMPUTATION PROCEDURES

Missing income values in *Understanding Society* are replaced using a combination of cross-sectional and longitudinal imputation methods. There are two steps to the *Understanding Society* imputation procedure. The first replaces missing values using cross sectional imputation methods (with some exceptions where we make use of longitudinal information (see carryover below)). A second step then replaces the first stage imputes using the longitudinal imputation method of Little and Su (1989). The overall approach is based on that used for the Australian household panel, (HILDA), described in Hayes and Watson (2009). The various imputation methods used are described in more detail below.

3.8.2.1. Cross sectional imputation methods

Cross-sectional imputation is carried out year by year through a range of parametric, semi-parametric and non-parametric methods. Parametric methods are: linear regression (for continuous variables), interval regression (for continuous censored variables), logistic regression (for binary variables), ordered logistic regression (for ordered variables), multinomial logistic regression (for non-ordered categorical variables). The semi-parametric and non-parametric methods are, respectively, predictive mean matching (PMM) and hot-deck imputation.

Parametric methods and PMM are used to impute wages, self-employment earnings, second job earnings, interests and dividends, plus their predictors for responding individuals. For responding individuals, wages, self-employment earnings, second job earnings, interests and dividends, and their predictors are imputed jointly using chained equations (ICE). Hot-deck is used to impute income sources for proxies and non-respondents (missing values in the variables defining the categories are set equal to their median).

All variables are imputed as reported except for wages and self-employment income, where we convert amounts reported net to gross where gross is not reported, using a deterministic model based on the tax and national insurance system. In computing total personal income, it is assumed that all other sources are reported gross, or are

not subject to taxation. Net income estimates are also included in the data set (see Section 3.8.6).

In what follows, we outline briefly the characteristics of the main cross-sectional imputation methods used.

Chained equations (ICE)

It is a multivariate imputation method used to impute a set of variables jointly. We used it to impute the main income variables for respondents plus their predictors. ICE allows for interdependence between income and auxiliary variables by considering univariate models estimated separately and sequentially through stochastic imputation, see van Buuren, Boshuizen et al. (1999) and Ragunathan, Lepkowski et al. (2001). This method has been already used in some major household panel surveys such as the ECHP.

The ICE starts by considering the following recursive (triangular) system of imputation equations,

$$\begin{cases} Y_1 = \alpha_{10} + X\beta_1 + u_1 \\ Y_2 = \alpha_{20} + X\beta_2 + \alpha_{21}Y_1 + u_2 \\ Y_3 = \alpha_{30} + X\beta_3 + \alpha_{31}Y_1 + \alpha_{32}Y_2 + u_3 \\ \vdots \\ Y_k = \alpha_{k0} + X\beta_k + \alpha_{k1}Y_1 + \alpha_{k2}Y_2 + \cdots + \alpha_{kk-1}Y_{k-1} + u_k \end{cases}$$

Here, Y1,Y2,…,Yk are the income and auxiliary variables to be imputed ordered from the one with the smallest percentage of missing values, Y1, to the one with the largest percentage of missing values Yk, X is a set of auxiliary variables observed for all individuals, α's and β's are parameters and u1,u2,…,uk are random errors. Such a recursive system allows us to carry out the imputation separately for each variable and sequentially. The sequential procedure is given by the following steps:

- estimation of the first equation and imputation of the missing values for Y1,
- estimation of the second equation using the imputed values to replace the missing values of Y1, and imputation of Y2,
- repetition of estimation and imputation steps sequentially for each of the following equations until when all k variables, Y1,Y2,…,Yk have been imputed.

We use stochastic imputation, that is, we draw the imputed values from the posterior predictive distribution of the variable to be imputed, conditional to the observed data. For more details about stochastic imputation we refer to Rubin (1987), Schafer (1997), and Kenward and Carpenter (2007).

This sequential estimation is consistent only if the recursive system is valid. Since this is not necessarily a valid assumption, ICE uses the imputed values produced using the above recursive system as starting values in an iterative imputation process. In other words, the starting values are used to begin a new cycle of imputations where each equation is estimated sequentially, but this time using as explanatory variables both X and all the imputed variables Y1,Y2,…,Yk excluding the

one used as dependent variable. At the end of this new cycle, a set of new imputed variables is produced and used to begin a further new cycle of imputations. These cycles of imputations are repeated until convergence. Notice that in practice some of the variables will exclude certain of the Xs and Ys variables in the imputations because it does not always make sense to use all variables as predictors.

Predictive mean matching (PMM)

It is used to impute benefits, pensions and other incomes. For a given variable, PMM replaces missing values with observed values from a donor, i.e. a respondent with non-missing information on the variable of interest. This is done in four steps i) regression models for the variable to be imputed are estimated ii) fitted values are produced iii) records with missing information (recipients) are matched to donors based on the fitted values computed in ii), iv) missing values are replaced with observed values from donors. See also Little (1988).

Hot-deck (HD)

For individuals with missing information, the hot-deck method identifies suitable donors within imputation classes. Characteristics reported in the data associated with the missing information are used to define imputation classes.

In *Understanding Society*, the hot-deck method is used to impute missing information for proxies and non-respondents. For proxies, where available, imputation classes are defined by reported bands on income and earnings and limited covariates. For non-respondents and proxies where income bands are missing, a richer set of covariates are used to define imputation classes including: employment and benefit carryovers, age, education, sex, ethnicity, housing tenure, marital status, durable good ownership, whether a parent and number of bedrooms. Once a suitable donor is identified, information on all income sources is carried over from the donor.

3.8.2.2. Longitudinal imputation methods

Methods for longitudinal imputation are used to take into account longitudinal patterns in the data. The longitudinal methods we use are the “Population Carryover” and the “Little and Su” methods.

Population Carryover (PC)

PC uses data from adjoining waves to replace missing wave information. With only one adjoining wave of non-missing data, the information is carried-over with probability one. When two waves of adjoining information are available, the information carried-over is chosen based on proportions reported in the non-missing population.

In *Understanding Society*, PC is used to impute employment status and benefit eligibility for non-respondents and proxies. These variables are then used as inputs into a hot-deck procedure (see above).

Little and Su (LS)

The LS method imputes missing values using a multiplicative model (see Little and Su 1989). The final imputation is the product of 3 terms: a trend effect (across waves), the recipient’s departure from the trend, and a residual effect donated from

another respondent with complete information for the corresponding income component.

In *Understanding Society*, a modified version of the LS method is implemented. When identifying a donor for the residual effect, rather than using the non-missing information only, we additionally make use of information imputed from the cross-sectional methods described above. In this way, the LS procedure forms the final step in the *Understanding Society* imputation process, where imputes from the other methods form inputs into the LS method. Missing income sources imputed as inapplicable using the cross-sectional methods do not receive a LS impute, neither do respondents applicable for an income source in only one wave.

3.8.3. ITEM NONRESPONSE ON INCOME VARIABLES IN THE INDIVIDUAL QUESTIONNAIRE

The imputation of earnings (wages and self-employment earnings from the first job and earnings from the second job) and investment income in the individual questionnaire is performed considering a separate equation for each of the income components.

With variables for which we have point information (a single value), we use either log linear or predictive mean matching models. For those variables where we have bracketed values rather than point information (for example in the case of dividends and interests) or when we have *a priori* information which allows us to bound the missing income variable, we use interval regression. The type of regression models used to impute missing explanatory variables depends on the level of measurement. That is, we use log linear regression for continuous variables, and binary, ordered and multinomial logit models, respectively, for dummy, ordinal and unordered categorical variables. The explanatory variables are a set of characteristics collected in the individual (personal) or household questionnaires. The specification of the models varies by income variable but it generally includes most of the following variables:

- personal socio-economic variables (age, sex, self-reported ethnic group, indicator for respondent born in the UK, marital status, education level, general health, current subjective financial situation);
- personal income variables (excluding the one used as the dependent variable);
- lagged income variables (just for Waves 2 and 3)
- household characteristics (number of children in the household, house tenure, house type, household size);
- job characteristics (log number of hours normally worked per week, log number of hours per months in a second job, log years of job tenure, permanent or temporary job, occupation (SOC 2000, 1 digit), number employed at the current job workplace (for employees), number of employees if self-employed, whether is self-employed and hires employees, whether the employment organization is private or not (only for employees), type of ownership if self-employed (sole ownership or partnership), an indicator for whether annual business accounts are prepared for the Inland Revenue for tax purposes if self-employed);
- household variables reflecting economic situation (log amount spent on food from food shops in four weeks prior to interview, log amount spent on

food eaten outside the home in four weeks prior to interview, log last year expenditure on domestic fuel (e.g. electricity and gas), number of bedrooms in the house, number of other rooms in the house, Council Tax band);
- Government Office Regions (GOR).

The imputation of the income sources in the income file (pensions and benefits) is performed in a second model where each income source is imputed using as predictors the other income sources of the income file, a set of demographics (age, age squared, number of children, number of children squared, sex, ethnicity, marital status, GOR), the income sources imputed in the previous stage (earnings for first and second job and investment income), and information on benefits and pensions in the previous year (total value and total number of benefits).

All variables are imputed as reported except for wages and self-employment income, where we convert amounts reported net to gross where gross is not reported, using a deterministic model based on the tax and national insurance system. In computing total personal income, it is assumed that all other sources are reported gross, or are not subject to taxation.

The imputation of the missing income sources in the individual questionnaire permits the computation of total earnings (i.e., the sum of income from the first and second job) and total income (the sum of earnings, plus investment income, pensions and benefits).for all adult non proxy respondents.

3.8.4. ITEM NONRESPONSE FOR INCOME VARIABLES IN THE PROXY QUESTIONNAIRE

The only income variables reported in the proxy questionnaires are the total gross earnings and total gross income. We impute missing values for these two variables again using hot-deck methods. The imputation is based on the sample of persons responding to the individual questionnaire together with the sample of individuals for whom a proxy questionnaire is available. Since, for non-proxies, total earnings cannot be higher than total income, we impose additional restrictions to the brackets such that this relationship also holds in the case of the imputed earnings and income for proxies. The explanatory variables used for the imputation of gross earning and gross income for proxies are those used in the imputation of the income variables for non-proxies and listed above.

3.8.5. INDIVIDUAL NON-RESPONDENTS WITHOUT PROXY QUESTIONNAIRE

For individual non-respondents in responding households with no proxy questionnaire, we impute total personal income using information from the household questionnaire only. The procedure used is again hot-deck. More precisely, whenever available, we use:

- individual socio-economic variables (age, sex, marital status, ethnicity, work)
- household socio-economic variables (household size, number of children in the household, whether there is nobody in the household who speaks English, whether the interview had to be translated, house type, an indicator for whether the person is owner of the house, the external condition of the address relative to the others, number of bedrooms in the house, number of other rooms in the house, value of the property for home owners, number of cars, number of durables, log last year's expenditure on domestic fuel (e.g.

electricity and gas), amount spent on food eaten outside the home in four weeks prior to interview, amount spent on food from food shops in four weeks prior to interview, weekly rent paid, whether the household can keep the accommodation warm enough)

- GOR and LDA.

3.8.6. NET INCOME ESTIMATES

The data also include estimates for income net of tax and national insurance. As far as possible we have tried to replicate the approach used by DWP in developing their Households Below Average Income (HBAI) estimates. There are however some deductions and some incomes sources which are not available to us from the questionnaire. It is hoped that a later release of the data will provide further is estimates of some of these other components. A technical working paper providing more information the derivation of these net income variables is in preparation.

Data are included at the household and individual level. The individual level, the total estimated net income is ***w*_netinc1**. At the household level, there are two variables: ***w*_hhnetinc1**, which is the sum of net incomes from all household members (including proxies and within household non-respondents), and ***w*_hhnetinc3**, which ***w*_hhnetinc1** less council tax liability.

Council tax benefit, which is included in ***w*_hhnetinc1** is netted off from the council tax liability and not included in total income.

In addition to these summary variables there are estimates of the different income components, following the structure used by HBAI. These are as follows:

Component 1: Labour income (***w*_inc1lab**)

w*_paynu_dv** – net usual pay (w*_inc1labem**)
w*_seearnnet_dv** – net self-employment income (w*_inc1labse**)
w*_jb2pay_dv** - gross pay in second job, less estimated tax and national insurance (w*_inc1labj2**)

Component 2: Miscellaneous income (***w*_inc2misc**)
Receipts reported in income record where ***w*_ficode** equals:

24 educational grant (not student loan or tuition fee loan
27 payments from a family member not living here
38 any other regular payment (not asked wave 1)

Component 3: private benefit income (***w*_inc3prben**)
Receipts reported in income record where ***w*_ficode** equals:

25 trade union / friendly society payment
26 maintenance or alimony
35 sickness and accident insurance

Component 5: investment income (***w*_inc5inv**)
***w*_fiyrinvinc_dv** annual income from savings and investments, divided by 12
Receipts reported in income record where ***w*_ficode** equals:

4 a private pension / annuity
28 rent from boarders or lodgers (not family members) living here
29 rent from any other property - less estimated tax

Component 6: pension income (***w***_**inc6pen**)
Receipts reported in income record where ***w***_**ficode** equals:

2 a pension from a previous employer
3 a pension from a spouse s previous employer

Component 7: social benefit income (***w***_**inc7sben**)
Receipts reported in income record where ***w***_**ficode** equals:

1 state retirement (old age) pension
5 a widow s or war widow s pension
6 a widowed mother s allowance / widowed parent s allowance
7 pension credit (includes guarantee credit & saving credit)
8 severe disablement allowance
9 industrial injury disablement allowance
10 disability living allowance
11 attendance allowance
12 carer s allowance (formerly invalid care allowance)
13 war disablement pension
14 incapacity benefit
15 income support
16 job seeker s allowance
18 child benefit (including lone-parent child benefit payments)
19 child tax credit
20 working tax credit (includes disabled person s tax credit)
21 maternity allowance
22 housing benefit
23 council tax benefit (offset against council tax)
30 foster allowance / guardian allowance
31 rent rebate (NI only)
32 rate rebate (NI only – offset against rates)
33 employment and support allowance
34 return to work credit
36 in-work credit for lone parents
37 other disability related benefit or payment
39 income from any other state benefit (not asked in wave 1)

Deduction component 9: council tax (***w***_**dep9ctax**)
Calculated from council tax band and local authority district (GB). Estimates for Northern Ireland rate charges have not been included.

The variables ***w***_**dep9ctax** and ***w***_**hhnetinc3** are only available in the Special Licence version of the data set.

3.9. EXAMPLE CODE FOR MATCHING FILES

We are including six examples of common data management tasks useful in analysing the data. Each task is illustrated with code for Stata. Because Stata is case-sensitive, we have not displayed file and variable names in upper case, but in lower case. To run these programmes on UKHLS data files, please remove the "_ip" suffix from the data file names. Statements beginning with // are comments.

The first task is distributing household level information to individual level. We can do this by merging household level file (such as ***w*_hhresp**) with an individual level file (such as ***w*_indresp**) within the same wave, see Figure 3.

Figure 3: Stata Code: Distributing household level information to individuals

```
// open the household level file
use a_hidp a_hhsize using a_hhresp_ip, clear

// sort it on the household identifier, w_hidp
sort a_hidp

// save this temporary file
save hhinfo, replace

// open the individual level file

use pidp a_hidp  a_marstat using a_indresp_ip, clear
// sort it on the household identifier, w_hidp
sort a_hidp

// merge it with the earlier saved file on w_hidp. The output shows how many
cases matched
merge m:1 a_hidp using hhinfo

// drop this variable – essential step
drop _merge
save final1, replace

// clean up unwanted files
erase hhinfo.dta
```

The second task is summarising individual level information at the household level. We will summarise individual level information within a household (number of 18-24 year olds in the household) and then match that onto the household level file, see Figure 4.

Figure 4: Stata Code: Summarizing individual level information to household level

```
// open household level file, keeping only relevant variables and saving
use a_hidp a_hhsize using a_hhresp
sort a_hidp
save hhinfo, replace

// open enumerated individual level file, keeping only relevant variables
use pidp a_hidp a_dvage using a_indall_ip, clear
                                          , clear
// create a variable that counts the number of 18-24                    ch household
bysort a_hidp: egen n1824= sum(a_dvage>=18 & a_dvage<=24)

// keep only first observation for every household
bysort a_hidp: keep if _n==1

// keep only household level information
keep a_hidp n1824

// now merging this household information with the household level file
sort a_hidp

// merge data files and save
merge 1:1 a_hidp using hhinfo, nogen
drop _merge
save final2, replace

// housekeeping
erase hhinfo.dta
```

In the fourth example we will match the information of wives onto that of their partners/spouses.

Figure 5: Stata Code: Matching individuals within a household

```
// Open the dataset with information on all persons in responding households and
// keep only those persons who have a spouse/partner in the household
use a_hidp a_pno a_hgpart a_sex a_dvage using a_indall_ip if a_hgpart>0, clear

// rename the prefix a_ to something that would indicate that this information relates
// to the spouse or partner
renpfix a_ sp_

// rename the spouse/partner pno variable to the respondent pno variable as this will
// be used to match on to the respondent information. Then sort and save the data
rename sp_hgpart a_pno
rename sp_hidp a_hidp
drop sp_pno
sort a_hidp a_pno
save spousepartner, replace

// Again open the data with information on all persons in responding households
use a_hidp a_pno a_hgpart a_sex a_dvage using a_indall_ip if a_hgpart>0, clear

// rename the prefix a_ to something that would indicate that this information relates
// to the respondent
renpfix a_ r_

// as we want to match on a_hidp and a_pno rename r_hidp and r_pno back to these
rename r_hidp a_hidp
rename r_pno a_pno

// Now sort and merge with the spouse partner file
sort a_hidp a_pno
merge 1:1 a_hidp a_pno using spousepartner
drop _merge
save final3, replace
erase spousepartner.dta
```

In the fifth example we will create a variable that measures the number of siblings in the household using the **egoalt** file. The resulting file can be merged with any individual level file.

Figure 6: Stata Code: Using the egoalt file to create household composition variables

```
// Load the data file that stores relationship information
use b_hidp b_epno b_relationship using b_egoalt_ip, clear

// create a variable that counts the number of siblings in the household
bysort b_hidp b_epno: egen nsiblings = sum(b_relationship>=14 &
b_relationship<=17)
lab var nsiblings "number of siblings in household"

// keep one observation per person
bysort b_hidp b_epno: keep if _n==1

// store data file
sort b_hidp b_epno
save final4, replace
```

To match individual level files across *two* waves into a long format do the following (for more waves add wave specific prefix in the foreach statement):

Figure 7: Stata Code: Merging individual files across waves into long format

```
foreach w in a b {
        // open the individual level file
        use pidp `w'_jbhas using `w'_indresp_ip, clear
        // drop the wave prefix from all variables
        renpfix `w'_
        // create a wave variable
        gen wave=strpos("ab", "`w'")
        // save one file for each wave
        save temp`w', replace
        }
// open the file for the first wave (wave a_)
use tempa, clear
foreach w in b {
        // append the files for second wave onwards
        append using temp`w'
        }
// save the long file
save final5, replace
// erase temporary files
foreach w in a b {
        erase temp`w'.dta
        }
```

To match individual level files across *two* waves into a wide format do the following (for more waves add wave specific prefix in the foreach statement):

Figure 8: Stata Code: Merging individual files across waves into wide format

```
use pidp a_jbhas using a_indresp_ip, clear
sort pidp
save temp, replace

foreach w in b {
        use pidp `w'_jbhas using `w'_indresp_ip, clear
        sort pidp
        merge 1:1 pidp using temp
        drop _merge
        sort pidp
        save temp, replace
}
save final6, replace
erase temp.dta
```

3.10. PRESERVING CONFIDENTIALITY

In preparing the data for the general release we have taken steps to maintain the confidentiality of responses. These include not releasing the full date of birth and not releasing the most detailed job-related SOC and SIC codes. Household income has been top coded. Open or narrative text, e.g., names of schools or employers, has not

been released since it may indirectly identify individuals. Geographical identifiers below the level of GORs are also not included in the general release.

Analysts may apply to gain access to restricted resources, see Section 4, below. The study has a Data Access Committee to take decisions on applications requesting access to restricted electronic data and biological samples from *Understanding Society*. Its aim is to allow important research to proceed while minimising risks, particularly to study participants.

4. DATA ACCESS

The data are released through the UK Data Service (UKDS) in SPSS and Stata formats. While documentation is released through the UKDS, we encourage users to consult the *Understanding Society* webpage. The documentation will develop over time. We have developed specific guides about major content areas such as the biomeasures or cognitive measures, and guides for issues that are frequently problematic for users such as selection of appropriate weights. We will continue to develop specific user guides over time.

4.1. RELEASE VERSIONS

4.1.1. END USER LICENCE (EUL) DATA

Most of the Wave 1- Wave 4 data has been released according to the conditions of the regular UK Data Services' End User License (EUL): http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx#/tab-end-user-licence. The data are listed as SN6614 -*Understanding Society*: Waves 1-4, 2009-2013.

In 2010-2012, *Understanding Society* augmented survey questions with direct health assessments and the collection of blood samples. Data from this collection are released under the standard EUL through the UKDS, SN7251 - *Understanding Society*: Waves 2-3 Nurse Health Assessment, 2010-2012.

4.1.2. SPECIAL LICENCE (SL) DATA

A number of sensitive data are released under Special Licence (SL). Researchers can apply for access to SL data through a UKDS application procedure, where they are required to justify their research objectives and explain why EUL data alone would be inadequate to reach those objectives. Researchers are asked to report publications resulting from using the data. The conditions for using SL data are provided at http://ukdataservice.ac.uk/get-data/how-to-access/conditions/special-licence.aspx

A SL version of the UKHLS Main Survey data that contains the month of birth, full occupational coding, full urban-rural classification, rare country of birth/nationality occurrences and uncapped income variables is available through the UKDS, SN6931- *Understanding Society*: Wave 1-4, 2009-2013: Special Licence Access.

Look-up tables containing the official school reference code for children in the UKHLS are released under SL, see SN7182.

Look-up tables linking *Understanding Society* survey data via the wave-specific household identifier with official geographical units are also released under SL. They include Local Authority Districts (SN6666), Census 2001 Area Classification for Output Areas (SN6674), Travel to Work Areas (SN6675), Westminister Parliamentary Constituencies (SN6668), Rural-urban Indicators (SN7454), Local Education Authorities (SN6671), and Primary Care organisations (SN6673). For further information about these geographical units see Office for National Statistics (ONS) Geography website, http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/index.html. The geographical datasets in the UKHLS are described in Rabe (2011) and in annual release notes. Note that we provide official geographical codes for Census 2001 but also for Census 2011 as they become available.

A SL data set linking Department for Transport Accessibility Statistics 2009-2011 at the LSOA level with the UKHLS household identifier is available through the UKDS, SN7533 - *Understanding Society*: Waves 1-3, 2009-2012: Special Licence Access, Geographical Accessibility. For further information on this file see Knies and Menon (2014).

4.1.3. SECURE ACCESS

Postcode grid references and full date of birth (SN 6676) can only be accessed under secure settings. Further information about data available under this access route and about how to apply for access can be found on the UKDS, see http://discover.ukdataservice.ac.uk/catalogue/?sn=6676&type=Data%20catalogue and the Secure Data Service, see http://ukdataservice.ac.uk/use-data/secure-lab.aspx.

4.2. REVISIONS TO PREVIOUS RELEASES

We release the preceding waves of data when we make a new edition available. Users should refer to the document “UKHLS 2014 Revisions”, see http://doc.ukdataservice.ac.uk/doc/6614/mrdoc/pdf/6614_ukhls_2014_revisions.pdf

We request that researchers using the data notify us about errors, inconsistencies, and other problems with the data identified during their use of the data. We make use of this information in improving the data.

Please raise any issues relating to data or data analysis with our User Support service at https://www.understandingsociety.ac.uk/support/projects/support.

We communicate information to members of the *Understanding Society* users group. Please register for the group using the registration form provided at the top right of the following page https://www.understandingsociety.ac.uk/support/projects/support.

The *Understanding Society* website has a Frequently Asked Questions (FAQ) https://www.understandingsociety.ac.uk/documentation/faq. The User Support forum also has a FAQ about data-related questions: https://www.understandingsociety.ac.uk/support/projects/support/wiki.

4.3. LINKS TO OTHER STUDIES IN THE UKHLS FAMILY

4.3.1. THE BRITISH HOUSEHOLD PANEL SURVEY

Data from the BHPS prior to joining the UKHLS can be obtained from the UK Data Service (SN5151 British Household Panel Survey, Waves 1-18, 1991-2009, http://www.esds.ac.uk/findingData/snDescription.asp?sn=5151). The study documentation is available at http://www.iser.essex.ac.uk/bhps.

4.3.2. THE WAVES 2-3 NURSE HEALTH ASSESSMENT

In 2010-2012, *Understanding Society* augmented survey questions with direct health assessments and the collection of blood samples, see McFall, Petersen et al. (2014). Data from the Wave 2 and Wave 3 health assessment are released through the UK Data Service (SN7251 *Understanding Society*: Waves 2-3 Nurse Health Assessment, 2010-2012, http://discover.ukdataservice.ac.uk/catalogue?sn=7251).

4.3.3. THE UKHLS INNOVATION PANEL

The *Understanding Society* project incorporates the UKHLS Innovation Panel (IP), a separate survey intended to support methodological research (https://www.understandingsociety.ac.uk/about/innovation-panel). Data from the IP has been released through the UK Data Service (SN6849 *Understanding Society:* Innovation Panel, Waves 1-5, 2008-2013, http://www.esds.ac.uk/findingData/snDescription.asp?sn=6849).

5. CITATIONS AND ACKNOWLEDGEMENTS

Any publication, whether printed, electronic or broadcast, based wholly or in part on the *Understanding Society* data collection provided by the UK Data Service must be accompanied by the correct citation and acknowledge the Institute for Social and Economic Research as the data provider and the UK Data Service as the data distributor. The acknowledgement, which gives credit to sponsors or distributors, is not a replacement for a proper citation.

5.1. CITATION OF THE DATA

The format for bibliographic references is as follows:

> University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society*: Wave 1-4, 2009-2013 [computer file]. 6th Edition. Colchester, Essex: UK Data Service [distributor], December 2014. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-6.

5.2. CITATION OF THE USER MANUAL

The User Manual is to be cited as follows:

> Knies, Gundi (ed.) (2014). *Understanding Society* –UK Household Longitudinal Study: Wave 1-4, 2009-2013, User Manual. Colchester: University of Essex.

5.3. ACKNOWLEDGMENTS

People who participated in writing sections of the documentation for this or prior releases include, in alphabetical order, Randy Banks, Michaela Benzeval, Nick Buck, Jon Burton, Paul Fisher, Laura Fumagalli, Olena Kaminska, Gundi Knies, Peter Lynn, Stephanie McFall, Alita Nandi and Jakob Petersen. Many people participated in preparing and processing the questionnaires and data. From the information technology side we recognize the contributions of Geoffrey Angel, Tom Butler, Jeannette Chin, Paul Groves Elaine Prentice-Lane, Paul Siddall and Catherine Yuen. From the survey research team, we recognize Noah Uhrig, Sarah Budd, Emily Dix, and Deborah Wiltshire. A small group was active in contributing code for derived variables and flagging issues in using the data. They include Gundi Knies, Jakob Petersen, Alita Nandi, Cara Booker, Mark Bryan, Alexandra Skew, and Mark Taylor.

6. REFERENCES

Berthoud, R., L. Fumagalli, et al. (2009). "Design of the *Understanding Society* Ethnic Minority Boost Sample." *Understanding Society* Working Paper **2009-02**.

Burton, J. and (Ed.) (2012). "*Understanding Society* Innovation Panel Wave 4: Results from Methodological Experiments." *Understanding Society* Working Paper **2012-06**.

Hayes, C. and H. Watson (2009). "HILDA Imputation Methods." HILDA Project Technical Paper Series, University of Melbourne **No 02/9**.

Kenward, M. and J. Carpenter (2007). "Multiple imputation: current perspectives." Statistical Methods in Medical Research **16**(3): 199-218.

Knies, G. and S. Menon (2014). "*Understanding Society*: Waves 1-3, 2009-2012: Special Licence Access, Geographical Accessibility. The UKHLS-Accessibility Data File User Guide " http://doc.ukdataservice.ac.uk/doc/7533/mrdoc/pdf/7533_ukhls_accessibility_userguide.pdf.

Little, R. J. A. (1988). "Missing Data Adjustments in Large Surveys." Journal of Business and Economic Statistics **6**(287-296).

Little, R. J. A. and H. L. Su (1989). Item Non-response in Panel Surveys. Panel Surveys. D. Kasprzyk, G. J. Duncan, G. Kalton and M. P. Singh. New York, Wiley.

Lynn, P. (2009). "Sample design for *Understanding Society*." *Understanding Society* Working Paper **2009-01**.

Lynn, P., J. Burton, et al. (2012). "An initial look at non-response and attrition." *Understanding Society* Working Paper **2012-02**.

McFall, S. (2013). "*Understanding Society*- UK Household Longitudinal Study: Cognitive Ability Measures." *Understanding Society* User Manual.

McFall, S., J. Petersen, et al. (2014). "*Understanding Society* Waves 2 and 3 Nurse Health Assessment, 2010-2012: Guide to Nurse Health Assessment." https://www.understandingsociety.ac.uk/documentation/health-assessment

Office for National Statistics (1991). Census: Key Statistics for Local Authorities. London, HMSO.

Office for National Statistics (2003). "Census 2001: Key Statistics for Local Authorities in England and Wales'." www.ons.gov.uk/ons/rel/census/census-2001-key-statistics/local-authorities-in-england-and-wales/.

Rabe, B. (2011). "Geographic Identifiers in *Understanding Society*." *Understanding Society* Working Paper **2011-01**.

Ragunathan, E. T., J. M. Lepkowski, et al. (2001). "A Multivariate technique for multiply imputing missing values using a sequence of regression models." Survey Methodology **27**(1): 85-95.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York, Wiley.

Schafer, J. (1997). Analysis of Incomplete Multivariate Data. London, Chapman & Hall.

Spanier, G. B. (1976). "Measuring dyadic adjustment: new scales for assessing the quality of marriage and similar dyads." Journal of Marriage and the Family **38**: 15-27.

Taylor, M. F. (2010). Weighting, imputation and sampling errors. British Household Panel Survey User Manual Volume A: Introduction, technical report and appendices. J. Brice, N. Buck and E. Prentice-Lane. Colchester, University of Essex**:** A5 1-13.

van Buuren, S., H. C. Boshuizen, et al. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis." Statistics in Medicine **18**: 681-694.